Aggregation of time series predictors, optimality in a locally stationary context

Andrés Sànchez Pérez

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2015-ENST-0051

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Andrés SANCHEZ PÉREZ

le 16 septembre 2015

Agrégation de prédicteurs pour des séries temporelles, optimalité dans un contexte localement stationnaire

Directeurs de thèse : **François ROUEFF**
Christophe GIRAUD

Jury

M. Patrice BERTAIL, Professeur, MODAL'X, Université Paris Ouest (Nanterre)	Membre
M. Olivier CATONI, Directeur de recherche, EXCESS, CNRS & CREST	Membre
M. Jérôme DEDECKER, Professeur, MAP5, Université Paris Descartes	Membre
M. Liudas GIRAITIS, Professeur, School of Economics and Finance, Queen Mary University of London	Rapporteur
M. Christophe GIRAUD, Professeur, LMO, Université Paris Sud (Orsay)	Directeur de thèse
M. François ROUEFF, Professeur, LTCI, Télécom ParisTech	Directeur de thèse
M. Gilles STOLTZ, Chargé de recherche, GREGHEC, CNRS & HEC Paris	Rapporteur

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Remerciements

Une thèse est une aventure aléatoire. À chaque instant, la façon dont elle se passe oscille entre la calamité et l'excellence. Une conclusion intéressante de ce travail est que la distribution de l'aventure aléatoire évolue au cours du temps. De nombreux ingrédients interviennent. J'ai eu la chance de pouvoir compter sur plusieurs qui ont contribué de façon positive à sa réalisation.

Être encadré par François et Christophe fut un privilège et un plaisir, je vous adresse mes premiers remerciements. Le moins qu'on puisse dire c'est que je n'aurais pas pu avoir de meilleurs directeurs. J'ai beaucoup apprécié nos discussions devant un tableau, votre disponibilité sans faille et la relation amicale qui nous lie. Vous m'avez donné pleine liberté pour ma recherche tout en me conduisant par le chemin de la rigueur scientifique. J'admire votre talent mathématique et votre capacité de travail.

Gilles, tu as aussi participé (parfois de façon anonyme, parfois de façon publique) à ma formation. L'écriture de mon premier article fut une école. Merci de tes commentaires pertinents et de tes rapports toujours très détaillés et constructifs.

Liudas, I am glad and honored to have you as reviewer in my jury. I have learned a lot from your work and from your very interesting remarks on my thesis.

Jérôme, Olivier et Patrice, merci à vous de vous être intéressés à ma recherche, d'avoir pris le temps de lire ce document et d'être membres de mon jury. Les discussions que nous avons eues lors de ma soutenance furent un beau couronnement de ces trois années de travail. Vos articles et livres, tout comme ceux de mes directeurs et de mes rapporteurs, ont été des références obligées dans mon parcours.

J'ai été financé par le Conseil Régional d'Île-de-France. Son programme Réseau de Recherche Doctoral en Mathématiques (RDM-IdF) m'a permis, en particulier, d'assister à de nombreuses conférences, colloques, ateliers et échanges avec des collègues au-delà des frontières parisiennes. Dans ce cadre, Dominique W. et Sylvie ont été mes interlocutrices privilégiées, toujours à l'écoute.

The most interesting travel I have made (from both, academic and cultural points of view) was my visit to Nir at Technion in Israel. Nir, you were a great host. Thank you for letting me explore your bandit world and also for your nice recommendations to start discovering your amazing country.

Cécile, Françoise et Julie, grâce aux efforts que vous avez déployés pendant deux ans, mon niveau de français a progressé et je vous en sais gré. Les pages de ce document qui sont écrites dans la langue de Molière ont bénéficié de la relecture et des conseils d'Aurélien S., de Laetitia et de Thomas M. Un grand merci à vous !

Dominique R., Florence et Laurence, merci à vous d'avoir facilité mes démarches administratives avec la DRH de Télécom, l'école doctorale et le département TSI.

Parallèlement à ma recherche, j'ai eu la chance d'enseigner. Pendant la première année je suis intervenu à plusieurs cours à l'ENS Cachan et à Télécom. J'ai collaboré avec et

beaucoup appris d'Eric M., de Gersende, de Pascal et aussi de François. Les deux années qui ont suivi m'ont mené à Centrale pour participer aux cours d'Erick et Gilles F. Je vous suis reconnaissant de la confiance que vous m'avez tous témoignée.
L'équipe STA du LTCI est un cadre exceptionnel pour mener une thèse. En plus des collègues déjà mentionnés… il y a tous les autres ! Florence d'A.-B., tes présentations toujours très claires et ton amabilité sont des exemples à suivre ! Olivier C., merci de tes références sur les méthodes numériques. Stéphan, merci de ton accueil chaleureux et de tes pistes précieuses pour ma recherche d'emploi. Anne et Joseph, nos échanges sur la façon de présenter un article, sur l'utilisation du cluster ou sur l'agrégation ont beaucoup aidé à que ce travail soit moins cryptique.
Sans discréditer nos conversations un peu plus sérieuses, on s'amuse bien entre les statisticiens juniors de Télécom ! Que ce soit au FIAP, pendant les pauses thé/café ou autour d'un gâteau aux carambars à la fin d'un séminaire, la bonne humeur est toujours au rendez-vous.
Une pensée toute particulière pour Cristina et Émilie K., vous avez bien veillé à notre cohésion en vous appuyant sur vos compétences culinaires. Merci pour tout ce que vous m'avez apporté et pour notre amitié.
Je suis très content d'avoir partagé mon quotidien avec Adel, Adrien, Alain, Albert, Alireza, Amandine, Anna, Antoine, Aurélien B., Beppe, Émilie C., Eric S., Eugène, Gemma, Giulio, Guillaume, Igor, Jean, Mael, Marc, Maxime S., Minh, Mony, Moussab, Nicolas G., Nicolas S., Olivier F., Olivier M., Paul I., Paul L., Polo, Romain, Ruocong, Sonia, Sylvain L. C., Sylvain R., Tahar, Yao et Yasir. Une mention spéciale pour Claire V., notre bandit girl 2014 - 2017, pour ton engagement dans l'organisation des évènements de la vie du labo ainsi que pour tes bons plans. Maxim, je tiens à te remercier de ton aide inestimable en programmation HTML, de nos discussions fructueuses et de ton esprit de camaraderie.
Je garde un très bon souvenir des rencontres que j'ai faites pendant mes déplacements professionnels. Benj, Carole, Clément, François P., Ilaria, Laure, Lucie, Mélisande, Nelo, Séb et Sarah en font partie. Thanks also to the welcoming international gang of Technion! Cristina M., Alessandra, Alexander, Ranieri and Esben, it was very nice to hang with you!
À partir de 2013 j'ai découvert une belle initiative associative dans laquelle je continue de m'impliquer. Ces remerciements ne seraient pas complets si je ne les adressais aussi à la team ASLIVE. Mes pensées vont à AnneSo G., AnneSo R.-B., Arnaud d. L., Arnaud S., Aurélien S., Auriane, Damien, David, Flo, Florent, Geo, Gwen, Hélène, Jackie, Laetitia, Marie, Maxime d. P., Mélanie, Moly, PA, Philippine, Solène, Thérèse, Thomas G., Toti, Ségo, Véro et pour tous nos protégés.
Mes autres amis "parisiens" et d'ailleurs, avec qui j'ai eu la joie de partager du bonheur, ont également marqué ces trois années. Adi, Ali, Catalina, Citlali, Daniel B., Ingmar, Jo et Nina, merci à vous pour nos galettes des rois, nos voyages et nos nombreuses sorties.
Claire F., je te remercie de ton soutien, des moments merveilleux que nous avons vécus ensemble et de l'amitié qui continue à nous lier. Thomas M., je tiens à te remercier de nos discussions quotidiennes, de nos repas conviviaux, et de ton amitié fidèle.
Agradezco también a aquellos amigos que han estado lejos físicamente pero siempre

presentes durante la realización de mi doctorado. Algunos fueron mis profesores, algunos compartieron conmigo una joven pasión por las matemáticas. Adriana, Alejandro, Arazoza, Bobby, Daniel D.-C., David, Enech, Ger, Javi, Karel, Katy, Liannet, Lili, Lisdy, Mario, Meilyn, Néstor, Pablo, Pachy, Pável, Peter, Rodri, Samper, Sergio, Teresita, Valia, Willy, Yane, Yos, a todos gracias. Tony, a ti en especial por mostrarme tus trabajos y por todo lo que tenemos en común.

Por último, y porque es lo más importante, gracias a mi familia. Su ejemplo siempre ha sido un motivo de inspiración para mí. Sin el empeño que pusieron en mi educación y en crearme las condiciones necesarias para cada etapa de mi vida, esta tesis no hubiese visto la luz. Me hizo muy feliz el tener presente a mis padres, Ana María y Andrés, y a mi hermana Anabel el día de mi defensa. Este trabajo se los dedico enteramente a ustedes.

Contents

1 Introduction en français

1.1 Contenu et notation

Le présent chapitre établit les bases de notre recherche. Dans la Section 1.2, nous présentons les modèles qui nous intéressent : certaines classes de processus faiblement dépendants et de processus localement stationnaires. Notre objectif final est de proposer des méthodes de prédiction efficaces pour ces modèles. La qualité d'une prédiction est mesurée par une fonction perte. Nous cherchons à ce qu'elle soit aussi faible que possible, généralement en espérance ou avec une forte probabilité. Ces notions sont formalisées dans un cadre général présenté dans la Section 1.3.1. Dans la Section 1.3.3 nous expliquons les caractéristiques d'optimalité associées aux algorithmes de prédiction que nous explorons. L'agrégation par poids exponentiels est la pierre angulaire de cette thèse, nous donnons un bref aperçu sur ce sujet dans la Section 1.4. La Section 1.5 contient une énumération des problèmes précis que nous abordons et la Section 1.6 une présentation de nos résultats principaux. La Section 1.7 propose des directions de recherche possibles dans la continuité de notre travail.

Tout au long de ce chapitre, pour $\boldsymbol{a} \in \mathbb{R}^q$ avec $q \in \mathbb{N}^*$, $\|\boldsymbol{a}\|$ dénote sa norme euclidienne, $\|\boldsymbol{a}\| = (\sum_{i=1}^q a_i^2)^{1/2}$ et $\|\boldsymbol{a}\|_1$ sa norme ℓ_1, $\|\boldsymbol{a}\|_1 = \sum_{i=1}^q |a_i|$. Les caractères gras représentent des vecteurs colonne et les caractères normaux leurs composantes, comme par exemple $\boldsymbol{y} = (y_i)_{i\in\mathbb{Z}}$. L'utilisation de sous-indexes avec deux points ':' fait référence à des composantes consécutives d'un vecteur $\boldsymbol{y}_{1:k} = [y_1 \ \dots \ y_k]'$ (vers l'avant), $\boldsymbol{y}_{k:1} = [y_k \ \dots \ y_1]'$ (vers l'arrière) ou à des éléments d'une suite $\boldsymbol{X}_{1:k} = [X_1 \ \dots \ X_k]'$ (vers l'avant), $\boldsymbol{X}_{k:1} = [X_k \ \dots \ X_1]'$ (vers l'arrière) ; dans tous les cas, il s'agit de vecteurs de dimension k.

1.2 Modèles

Les variables aléatoires indépendantes et identiquement distribuées sont la matière première d'une ample partie de la littérature statistique. Bien que la présente contribution repose sur ces variables-là, elles ne sont pas notre cible principale. Les problèmes particuliers que nous étudions se concentrent sur des suites de variables aléatoires qui peuvent être (et il est intéressant quand elles le sont) dépendantes et qui peuvent avoir (et il est intéressant quand elles en ont) une distribution qui évolue. Les deux prochains sous sections mettent brièvement les modèles que nous étudions en contexte.

1.2.1 Dépendance faible

Le paradigme de dépendance faible, proposé par Doukhan and Louhichi (1999), est une approche qui rend explicite l'indépendance asymptotique entre deux moments distants d'une série temporelle. Il représente un point de vue unificateur d'autres notions concurrentielles telles que les conditions de mélange, plus adaptées à des σ-algèbres. Les coefficients de mélange α (fort) et β par exemple, ont été introduits par Rosenblatt (1956) et Volkonskiĭ and Rozanov (1959), respectivement. Nous revisitons quelques définitions qui sont essentielles dans notre recherche. Nous nous référons Dedecker et al. (2007) comme littérature très complète dans l'étude de la dépendance faible. Rappelons que deux variables aléatoires X et Y définies sur le même espace de probabilité sont dites indépendantes si et seulement si $\mathrm{cov}(f(X), g(Y)) = 0$ pour toutes les fonctions f et g mesurables et bornées. Un relâchement de la condition d'indépendance suit.

Définition 1 (Dépendance faible)**.** *La suite $(X_t)_{t\in\mathbb{Z}}$ à valeurs dans un espace topologique localement compact $\mathcal{X}$ (typiquement $\mathbb{R}^d$) est dite faiblement dépendante, s'il existe une classe $\mathcal{F}$ de fonctions telle que pour tout $u, v \in \mathbb{N}^*$ et toutes $f, g \in \mathcal{F}$ respectivement définies sur $\mathcal{X}^u$ et $\mathcal{X}^v$ la relation asymptotique suivante est valable*

$$\varepsilon(r) = \sup_{i_1\leq\ldots\leq i_u<i_u+r\leq j_1\leq\ldots\leq j_v} \left|\mathrm{cov}\left(f\left(X_{i_1},\ldots,X_{i_u}\right), g\left(X_{j_1},\ldots,X_{j_v}\right)\right)\right| \to 0\,, \text{ quand } r\to\infty\,.$$

Les décalages de Bernoulli sont une classe très riche de processus faiblement dépendants. C'est le premier modèle que nous étudions.

Définition 2 (Décalages de Bernoulli)**.** *Soit $(\xi_t)_{t\in\mathbb{Z}}$ une suite de variables aléatoires réelles indépendantes et soit $H : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ une fonction borélienne. Un décalage de Bernoulli est une suite $(X_t)_{t\in\mathbb{Z}}$ satisfaisant*

$$X_t = H\left(\xi_{t-j}, j\in\mathbb{Z}\right)\,. \tag{1.2.1}$$

Pas toutes les suites $(\xi_t)_{t\in\mathbb{Z}}$ et fonctions mesurables $H : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ définissent un décalage de Bernoulli. Un souci de convergence peut émaner de l'expression de H et des particularités de ξ_t. Afin d'illustrer cette ambigüité, considérons ξ_t une variable aléatoire uniforme en $[1, 2]$ pour tout t et $H(\boldsymbol{u}) = \sum_{j\in\mathbb{Z}}(-1)^j u_j$. Dans ce cas, la représentation (1.2.1) est dépourvue de sens.
L'expression (1.2.1) est bien définie lorsque les $(\xi_t)_{t\in\mathbb{Z}}$ ont des moments absolus uniformément bornés et que H est Lipschitz, c'est-à-dire si $\sup_{t\in\mathbb{Z}}\mathbb{E}[|\xi_t|] < \infty$ et pour tout $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathbb{Z}}$

$$|H(\boldsymbol{u}) - H(\boldsymbol{v})| \leq \sum_{j\in\mathbb{Z}} A_j \left|u_j - v_j\right|\,, \tag{1.2.2}$$

avec

$$A_* = \sum_{j\in\mathbb{Z}} A_j < \infty\,. \tag{1.2.3}$$

Si nous considérons $\mathcal{F} = \cup_{j\geq 1}\mathcal{F}_j$ dans la Définition 1, où $\mathcal{F}_j$ est l'ensemble de fonctions Lipschitz bornées de $\mathbb{R}^j$ vers $\mathbb{R}$, nous pouvons montrer que le précédemment bien défini décalage de Bernoulli (qui satisfait (1.2.1), (1.2.2) et (1.2.3)) est faiblement dépendant (voir (Doukhan and Louhichi, 1999, Lemma 9) et (Dedecker et al., 2007, Lemma 3.1)).
Si pour tout $t \in \mathbb{Z}$, l'instance X_t dépend uniquement de $(\xi_s)_{s\leq t}$, à savoir

$$X_t = H\left(\xi_{t-j}, j \geq 0\right)\,,$$

nous disons que le processus $(X_t)_{t\in\mathbb{Z}}$ est un décalage de Bernoulli causal (CBS en anglais) et les variables aléatoires $(\xi_t)_{t\in\mathbb{Z}}$ sont appelées innovations.
Les décalages de Bernoulli regroupent plusieurs processus faiblement dépendants dérivés de suites stationnaires. Ils fournissent aussi des exemples de processus faiblement dépendants mais pas mélangeants (voir Rosenblatt (1980)). Dans la suite nous présentons deux exemples de décalage de Bernoulli.

Exemple 1 (Processus à moyenne glissante infinie (MA(∞))). Soit $(\xi_t)_{t\in\mathbb{Z}}$ une suite de variables aléatoires i.i.d., centrées et avec variance 1. Le processus MA(∞) est défini par la représentation

$$X_t = \sum_{j\in\mathbb{Z}} a_j\xi_{t-j}\,,$$

où $\sum_{j\in\mathbb{Z}} |a_j| < \infty$.

Les processus de Volterra sont une généralisation des MA(∞) de l'Exemple 1 (voir (Doukhan, 2003, Section 2.4)).

Exemple 2 (Processus de Volterra). Soit $(\xi_t)_{t\in\mathbb{Z}}$ une suite de variables aléatoires i.i.d., centrées et avec variance 1 et soit $v_0 \in \mathbb{R}$. Nous considérons la suite $(a_{k;i_1,\ldots,i_k})_{k\in\mathbb{N}^*,(i_1,\ldots,i_k)\in\mathbb{Z}^k}$ de nombres réels. Posons

$$\begin{aligned} X_t &= v_0 + \sum_{k=1}^{\infty} V_{k,t}\,, \\ V_{k,t} &= \sum_{i_1<\ldots<i_k} a_{k;i_1,\ldots,i_k} \prod_{j=1}^{k} \xi_{t-i_j}\,. \end{aligned}$$

Observez que si les coefficients satisfont

$$\sum_{k=1}^{\infty} \sum_{i_1<\ldots<i_k} \left|a_{k;i_1,\ldots,i_k}\right| < \infty\,,$$

alors, l'inégalité de Minkowski implique que $X_t \in L^2$ pour tout $t \in \mathbb{Z}$.

Nous clôturons la section présente en donnant un résultat utile.

Une inégalité de concentration :

Dans les processus faiblement dépendants, à l'instar du cadre i.i.d. qui est plus classique, les preuves des résultats concernant la qualité de la prédiction impliquent l'utilisation d'inégalités de concentration (voir Massart (2007)). Par souci de clarté, nous présentons une inégalité exponentielle de type Hoeffding (voir (Rio, 2000, Theorem 1) et (Alquier and Wintenberger, 2012, Proposition 4.2)) satisfaite par les CBS.

Soit $n > 0$. Nous disons que $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction M-Lipschitz si pour tout $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq M \|\boldsymbol{x} - \boldsymbol{y}\|_1 \ .$$

Theorem 1.2.1. *Soit $(X_t)_{t\in\mathbb{Z}}$ un CBS borné associé à des innovations bornées $(\xi_t)_{t\in\mathbb{Z}}$, soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction 1-Lipschitz et $x > 0$. L'inégalité suivante est satisfaite*

$$\mathbb{P}\left(f(\boldsymbol{X}_{1:n}) - \mathbb{E}\left[f(\boldsymbol{X}_{1:n})\right] \geq x\right) \leq \exp\left(-x^2\left[2n\left(\|X_0\|_\infty + 2\sum_{j\geq 0} jA_j \|\xi_0\|_\infty\right)\right]^{-1}\right),$$

où la suite $(A_j)_{j\geq 0}$ satisfait l'Inégalité (1.2.2), et $\|X_0\|_\infty$ et $\|\xi_0\|_\infty$ sont les valeurs suprêmes respectives de $|X_0|$ et $|\xi_0|$.

1.2.2 Stationnarité locale

Une propriété particulière des processus stationnaires est qu'ils impliquent une collection de variables aléatoires identiquement distribuées. Une classe moins restrictive est celle des processus stationnaires au sens faible (ou de second ordre, ou en covariance). Nous rappelons la définition (voir (Brockwell and Davis, 2002, Section 1.4 et Theorem 2.1.1)).

Définition 3 (Processus faiblement stationnaires)**.** *Soit $\mu \in \mathbb{R}$ et $\gamma : \mathbb{Z} \to \mathbb{R}$ une fonction symétrique et définie non-négative. Nous disons que le processus à valeurs réelles $(X_t)_{t\in Z}$ est faiblement stationnaire si les trois conditions suivantes sont satisfaites.*

(i) $\mathbb{E}|X_t|^2 < \infty$ *pour tout* $t \in \mathbb{Z}$,

(ii) $\mathbb{E}[X_t] = \mu$ *pour tout* $t \in \mathbb{Z}$,

(iii) $\mathrm{Cov}(X_s, X_t) = \gamma(s-t)$ *pour tout* $s, t \in \mathbb{Z}$.

Une caractéristique cruciale des processus faiblement stationnaires est que leur spectre est constant. Plusieurs études portant sur des séries temporelles à spectre évolutif (parmi d'autres caractéristiques) sont apparues dans la seconde moitié du siècle dernier (voir par exemple Granger (1964) et Priestley (1965)). Trois décennies plus tard, Dahlhaus (1996b) introduisit une approche permettant une féconde analyse asymptotique locale.
Supposons par exemple que les observations correspondent au modèle

$$X_t = \theta_t X_{t-1} + \sigma_t \xi_t \ . \tag{1.2.4}$$

Au premier regard, nous pouvons nous demander comment un estimateur de la fonction θ, obtenu à partir de l'échantillon $(X_t)_{1\leq t\leq T}$, se comporte quand T est assez grand. Considérer ce type de question est usuellement contradictoire avec la supposition de non-stationnarité. Étant donné que la structure probabiliste du processus peut substantiellement différer des plus petites aux plus grandes valeurs de t, l'information qui arrive avec des nouvelles observations (disons t assez grand) peut être inutile pour estimer θ_t pour des valeurs petites de t.
Afin de surmonter cette difficulté, Dahlhaus (1996b) proposa l'idée des processus localement stationnaires, lesquels admettent la représentation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} \exp\left(it\omega\right) A^0_{t,T}\left(\omega\right) \xi\left(\mathrm{d}\omega\right) \ , \tag{1.2.5}$$

où, en particulier, ils existent une constante K et une (unique) fonction $2\pi-$ périodique $A : (-\infty, 1] \times \mathbb{R} \to \mathbb{C}$ avec $A(u, -\omega) = \overline{A(u, \omega)}$ tel que pour tout T

$$\sup_{t,\omega} \left| A^0_{t,T}\left(\omega\right) - A\left(\frac{t}{T}, \omega\right) \right| \leq \frac{K}{T} \ . \tag{1.2.6}$$

L'introduction artificielle de la dépendance en l'horizon T et des hypothèses supplémentaires sur $(X_t)_{1\leq t\leq T}$, ouvrent la voie à des procédures statistiques asymptotiques qui ont du sens. Ce modèle de séries localement stationnaires couvre essentiellement les processus linéaires variables dans le temps. Une partie considérable de notre travail est en lien avec lui.
Dans le cas du processus décrit par l'Équation (1.2.4), par exemple, il est localement stationnaire quand la suite $(\theta_{t,T})_{1\leq t\leq T}$ satisfait

$$\sup_{T\geq 1} \sum_{t=1}^{T} \left| \theta_{t,T} - \theta\left(\frac{t}{T}\right) \right| < \infty \ , \tag{1.2.7}$$

où $\theta : [0, 1] \to \mathbb{R}$ est une fonction adéquate (voir Dahlhaus (2009) et les références qui s'y trouvent).

Exemple 3 (Modèle TVAR). Une version particulière du modèle (1.2.4) est le processus auto-régressif variable dans le temps (TVAR en anglais). Il satisfait l'équation récursive suivante

$$X_{t,T} = \sum_{j=1}^{d} \theta_j \left(\frac{t}{T}\right) X_{t-j,T} + \sigma\left(\frac{t}{T}\right)\xi_t \ ,$$

où $(\xi_t)_{t\in\mathbb{Z}}$ est un bruit blanc et $\boldsymbol{\theta} = [\theta_1 \ \ldots \ \theta_d] \in s_d(\delta)$ avec $\delta \in (0,1)$.
L'ensemble $s_d(\delta)$ est lié à la stabilité de $(X_{t,T})_{1\le t\le T}$ et se définit par

$$s_d(\delta) = \left\{ \boldsymbol{\theta} : (-\infty, 1] \to \mathbb{R}^d, 1 - \sum_{j=1}^{d} \theta_j(u) z^j \neq 0, \forall\, |z| < \delta^{-1}, u \in [0,1] \right\} \ , \quad (1.2.8)$$

(voir (Dahlhaus, 1996b, Theorem 2.3)).
Des conditions de régularité sur $\boldsymbol{\theta}$ sont nécessaires afin d'obtenir des résultats intéressants pour les processus TVAR. Exiger des dérivées jusqu'à un certain ordre en est une assez standard (voir (Dahlhaus and Giraitis, 1998, Assumption 2.1 (i) et Assumption 3.1 (ii)-(iii))). Moulines et al. (2005) se repose sur une hypothèse plus flexible comparée à Dahlhaus and Giraitis (1998) : soit $R, \beta > 0$ et soit k le plus grand entier strictement plus petit que β ; en plus de $\boldsymbol{\theta} \in s_d(\delta)$, ils supposent que $\boldsymbol{\theta} \in \Lambda_d(\beta, R)$, où

$$\Lambda_d(\beta, R) = \left\{ \boldsymbol{\theta} \in C^k\left((-\infty, 1], \mathbb{R}^d\right) : \sup_{0<|s-s'|<1} \frac{\left|\boldsymbol{\theta}^{(k)}(s) - \boldsymbol{\theta}^{(k)}(s')\right|}{|s-s'|^{\beta-k}} \le R \right\} \ . \quad (1.2.9)$$

Des idées similaires ont été développées dans des contextes différents. Les processus auto-régressifs conditionnellement hétéroscédastiques variables dans le temps (tvARCH en anglais) (voir (Dahlhaus and Subba Rao, 2006, Section 2)) et les processus auto-régressifs conditionnellement hétéroscédastiques généralisés variables dans le temps (tvGARCH en anglais) (see (Subba Rao, 2006, Section 5)) sont des exemples qu'on peut citer. Un ingrédient commun à toutes ces approches est que le processus peut être localement approximé par sa version stationnaire.
Dans la suite nous introduisons une extension simple du modèle linéaire localement stationnaire. Soit $(Z_t)_{t\in\mathbb{Z}}$ une suite de variables aléatoires non négatives (pas nécessairement i.i.d.). Le processus $(X_t)_{t\in\mathbb{Z}}$ est dit sous-linéaire par rapport à $(Z_t)_{t\in\mathbb{Z}}$ si

$$|X_t| \le \sum_{j\in\mathbb{Z}} A_t(j) Z_{t-j} \ , \quad (1.2.10)$$

où $(A_t(j))_{t,j\in\mathbb{Z}}$ sont des coefficients non-négatifs qui satisfont

$$A_* := \sup_{t\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} A_t(j) < \infty \ .$$

Des résultats concernant $(X_t)_{t\in\mathbb{Z}}$ sont déduits en imposant des hypothèses additionnelles sur $(Z_t)_{t\in\mathbb{Z}}$. Par exemple, l'inégalité de Minkowski implique l'existence de moments d'ordre p pour le processus $(X_t)_{t\in\mathbb{Z}}$ quand les moments d'ordre p of $(Z_t)_{t\in\mathbb{Z}}$ sont uniformément bornés.

Exemple 4 (Un modèle non-linéaire). Un exemple de processus sous-linéaire et en même temps non-linéaire est donné par

$$X_t = g_t(X_{t-1}) + \xi_t \ ,$$

où $(\xi_t)_{t\in\mathbb{Z}}$ sont i.i.d. et $(g_t)_{t\in\mathbb{Z}}$ est une suite variable dans le temps de fonctions sous-linéaires qui satisfont, pour tout t

$$|g_t(x)| \leq \alpha|x| \ ,$$

pour un $\alpha \in (0,1)$. Alors, nous avons que

$$|X_t| \leq \alpha|X_{t-1}| + |\xi_t| \ .$$

En itérant cette équation vers l'arrière nous obtenons (1.2.10) avec $Z_t = |\xi_t|$ et $A_t(j) = \alpha^j$. Dans un cadre stationnaire, où $g = g_t$ est indépendante de t, une illustration bien connue de ce cas non-linéaire est donnée par le modèle auto-régressif seuillé où g est linéaire par parties, voir Tong and Lim (1980).

1.3 PRÉDICTION

1.3.1 Cadre général

Dans cette section nous présentons un cadre général qui inclut un nombre important de problèmes de prédiction étudiés dans la littérature. Nous considérons une série temporelle $(Z_t)_{1\leq t\leq T}$. La construction de prédicteurs à un pas $(\widehat{Z}_t)_{1\leq t\leq T}$ repose parfois sur un ensemble de données d'apprentissage. Un cas typique est quand on divise les données en un ensemble d'apprentissage et un ensemble de validation. Les prédicteurs apprennent exclusivement de l'ensemble d'apprentissage, tandis que l'ensemble de validation est utilisé pour évaluer la qualité de la prédiction. Dans la suite, nous fournissons les notions requises afin de construire notre formalisme.
Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilités, soit $\mathcal{H} \subset \mathcal{F}$ une sous σ-algèbre que nous appelons la σ-algèbre d'apprentissage, et soit $(\mathcal{F}_t)_{t\geq 0}$ une filtration que nous appelons la filtration de prédiction. La σ-algèbre $\mathcal{H}$ contient l'information de l'ensemble d'apprentissage. Elle peut être réduite à la σ-algèbre triviale si aucun ensemble d'apprentissage n'est disponible.
Considérez un processus $(Z_t)_{t\geq 0}$ à valeurs dans $\mathcal{Z}$, adapté à $(\mathcal{F}_t)_{t\geq 0}$, où $(\mathcal{Z}, \ell)$ est un espace métrique.

Définition 4 (Prédicteur)**.** *Pour tout $t \geq 1$, nous disons que $\widehat{Z}_t$ est un prédicteur de Z_t s'il est mesurable par rapport à la σ-algèbre $\mathcal{H} \vee \mathcal{F}_{t-1}$.*
Pour tout $T \geq 1$, nous dénotons par $\mathcal{P}_T$ la famille de suites $\widehat{Z} = (\widehat{Z}_t)_{1 \leq t \leq T}$ de prédicteurs de $(Z_t)_{1 \leq t \leq T}$, c'est-à-dire l'ensemble de tous les processus $\widehat{Z} = (\widehat{Z}_t)_{1 \leq t \leq T}$ adaptés à $(\mathcal{H} \vee \mathcal{F}_{t-1})_{1 \leq t \leq T}$.

Nous définissons la perte d'apprentissage comme

$$\frac{1}{T} \sum_{t=1}^{T} \ell\left(\widehat{Z}_t, Z_t\right) . \tag{1.3.1}$$

Le risque de prédiction est fourni par l'espérance conditionnelle de la perte d'apprentissage étant donnée la σ-algèbre d'apprentissage.

$$R_T\left(\widehat{Z} \middle| \mathcal{H}\right) = \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \ell\left(\widehat{Z}_t, Z_t\right) \middle| \mathcal{H} \right] . \tag{1.3.2}$$

Le risque est défini comme l'espérance de la perte d'apprentissage.

$$R_T\left(\widehat{Z}\right) = \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \ell\left(\widehat{Z}_t, Z_t\right) \right] . \tag{1.3.3}$$

Il convient d'avoir en tête que, en fonction du contexte, une suite de prédicteurs $\widehat{Z}$ est plus efficace quand (1.3.1), (1.3.2) ou (1.3.3) sont plus petits. Souvent, quand on divise nos données, on cherche à minimiser (1.3.2) avec une grande probabilité. D'un autre côté, si $\mathcal{H}$ est la σ-algèbre triviale, (1.3.2) et (1.3.3) coïncident et on cherche à les minimiser.
Les deux sections qui suivent expliquent comment des problèmes classiques de prédiction liés à cette thèse rentrent dans le cadre décrit. Elles fournissent aussi des résultats standards.

1.3.1.1 Sur la prédiction d'un processus dépendant étant donné un ensemble d'apprentissage

Supposons qu'on observe les premières T instances d'un processus stochastique $X = (X_t)_{t \geq 1}$ possiblement dépendant à valeurs dans $\mathcal{X}$. La distribution du processus entier est dénotée par P. Supposons en plus que nous souhaitons prédire les premières T instances d'un autre processus stochastique $Y = (Y_t)_{t \geq 1}$, indépendant de X, à valeurs dans $\mathcal{X}$ et distribué également selon P.
Ce contexte se présente souvent à nous quand on divise les données en deux ensembles : apprentissage et validation. Nous nous référons à Audibert and Catoni (2010, 2011); Hsu et al. (2011) quand toutes les observations sont indépendantes. Une situation plus complexe se produit quand les données disponibles sont dépendantes (à l'instar d'un processus AR(d)). Même si la prémisse de l'indépendance n'est pas remplie, dans la

pratique, nous pouvons diviser les observations et espérer que les prédicteurs construits à partir de l'ensemble d'apprentissage (X) conduisent à un risque faible dans l'ensemble de validation (Y). Pour des propos théoriques, il est convenable de supposer que X et Y sont indépendants, bien que cela puisse ne pas être vrai dans la pratique.
Soit $\mathcal{H} = \sigma(X_t, 1 \leq t \leq T)$ la σ-algèbre d'apprentissage et $\mathcal{F} = (\mathcal{F}_t)_{t\geq 1}$ la filtration naturelle associée à Y, où $\mathcal{F}_t = \sigma(Y_s, 1 \leq s \leq t)$. Dans ce contexte, nous construisons pour chaque $t = 1, \dots, T$ une application $\hat{f}_t : \mathcal{X}^T \to \mathcal{X}^{\mathcal{X}^{t-1}}$. Nous dénotons par $\hat{f}_t(\cdot|\boldsymbol{X}_{1:T})$ la fonction qui prédit Y_t en utilisant $Y_1, \dots, Y_{t-1}$, étant données les observations $X_1, \dots, X_T$. Alors, faisons $\widehat{Y}_t = \hat{f}_t(\boldsymbol{Y}_{1:t-1}|\boldsymbol{X}_{1:T})$.
L'expression suivante correspond au risque de prédiction défini par l'Équation (1.3.2)

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell\left(\hat{f}_t\left(\boldsymbol{Y}_{1:t-1}|\,\boldsymbol{X}_{1:T}\right), Y_t\right)\middle|\,\boldsymbol{X}_{1:T}\right] = \frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{X}^{\mathbb{N}^*}}\ell\left(\hat{f}_t\left(\boldsymbol{y}_{1:t-1}|\boldsymbol{X}_{1:T}\right), y_t\right)P\left(\mathrm{d}\boldsymbol{y}\right) \ . \quad (1.3.4)$$

Considérons le cas où le processus est stationnaire et le prédicteur $\hat{f}_t(\cdot|\boldsymbol{X}_{1:T})$ exploite uniquement $Y_{t-d}, \dots, Y_{t-1}$ pour un $d > 0$ fixé. Par ailleurs, nous admettons que $\hat{f}_t$ ne dépend pas de t. Ôter les premiers d termes de (2.3.4) nous conduit à analyser

$$R\left(\widehat{Y}\middle|\mathcal{H}\right) = \int_{\mathcal{X}^{\mathbb{N}^*}}\ell\left(\hat{f}\left(\boldsymbol{y}_{1:d}|\boldsymbol{X}_{1:T}\right), y_{d+1}\right)P\left(\mathrm{d}\boldsymbol{y}\right) \ ,$$

qui ne dépend plus de T.
Le risque correspondant prend la forme

$$\mathbb{E}\left[\ell\left(\hat{f}\left(\boldsymbol{Y}_{1:d}|\boldsymbol{X}_{1:T}\right), Y_{d+1}\right)\right] = \int_{\mathcal{X}^{\mathbb{N}^*}}\int_{\mathcal{X}^{\mathbb{N}^*}}\ell\left(\hat{f}\left(\boldsymbol{y}_{1:d}|\boldsymbol{x}_{1:T}\right), y_{d+1}\right)P\left(\mathrm{d}\boldsymbol{y}\right)P\left(\mathrm{d}\boldsymbol{x}\right) \ .$$

Exemple 5 (AR(∞)). Considérez les observations $X_1, \dots, X_T$, d'un processus auto-régressif $X = (X_t)_{t\in\mathbb{Z}}$. Il est défini par l'équation récursive

$$X_t = \sum_{j=1}^{\infty}\theta_j X_{t-j} + \sigma\xi_t \ , \quad (1.3.5)$$

où $\boldsymbol{\theta} = [\theta_1 \ \dots]' \in \mathbb{R}^{\mathbb{N}^*}$, $|\boldsymbol{\theta}|_1 = \sum_{j=1}^{\infty}|\theta_j| < \infty$, $1 - \sum_{j=1}^{\infty}\theta_j z^j \neq 0$ pour $|z| \leq 1$, $\sigma > 0$, et $(\xi_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires i.i.d. centrées et de variance 1.

Prédiction linéaire pour un AR(∞) :
L'approche standard pour prédire cette classe de processus date de Akaike (1969) (voir des extensions et généralisations dans Berk (1974), Bhansali (1978) et Lewis and Reinsel (1985)). Elle consiste à, pour un $d \in \mathbb{N}^*$ fixé, régresser X_t dans $X_{t-1}, \dots, X_{t-d}$. En d'autres termes, nous construisons un prédicteur $\widehat{X}_t = \widehat{\boldsymbol{\theta}}'\boldsymbol{X}_{t-1:t-d}$, où $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1 \ \dots \ \widehat{\theta}_d]' \in \mathbb{R}^d$.

Étant donné $d \in \mathbb{N}^*$, l'estimateur $\widehat{\boldsymbol{\theta}}$ est supposé réaliser l'infimum de l'erreur quadratique moyenne en prédiction

$$\frac{1}{T}\sum_{t=1}^{T}\left(X_t - \widehat{\boldsymbol{\theta}}'\boldsymbol{X}_{t-1:t-d}\right)^2 .$$

Par conséquent, elle est donnée par les équations Yule-Walker

$$\widehat{\boldsymbol{\theta}} \;=\; \widehat{\Gamma}^{-1}\widehat{\boldsymbol{\gamma}} ,$$

où $\widehat{\boldsymbol{\gamma}} = [\widehat{\gamma}_1 \;\ldots\; \widehat{\gamma}_d]'$, $\widehat{\Gamma}$ est la matrice de covariances empiriques $\widehat{\Gamma} = (\widehat{\gamma}_{i-j};\, i,j = 1,\ldots,d)$, qui doit être inversible, et $\widehat{\gamma}$ est fonction de covariance empirique

$$\widehat{\gamma}_\ell = \frac{1}{T}\sum_{t=1}^{T-|\ell|} X_t X_{t+|\ell|} .$$

L'expression précédente suppose que le processus est centré, c'est-à-dire $\mathbb{E}[X_t] = 0$ pour tout t. Considérez maintenant $Y = (Y_t)_{t\geq 1}$, une copie indépendante de X et soit $\hat{f}(\boldsymbol{y}_{1:d}|\boldsymbol{X}_{1:T}) = \widehat{\boldsymbol{\theta}}'\boldsymbol{y}_{d:1}$, où $\widehat{\boldsymbol{\theta}}$ est calculé à partir de X. Sous des hypothèses assez modérées, qui incluent l'existence de moments d'ordre 4 pour ξ_t (voir les détails en (i)-(iv) du (Bhansali, 1978, Theorem 1)), le résultat asymptotique suivant est établi (voir (Bhansali, 1978, Equation (4.5))) pour d et T assez grands

$$\mathbb{E}\left[\left(\hat{f}\left(\boldsymbol{Y}_{1:d}\,|\boldsymbol{X}_{1:T}\right) - Y_{d+1}\right)^2\right] - \sigma_d^2 \sim M\frac{d}{T} , \tag{1.3.6}$$

où $M > 0$ et $\sigma_d^2 = \inf_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathbb{E}[(\boldsymbol{\theta}'\boldsymbol{Y}_{d:1} - Y_{d+1})^2]$.

Choix de d :

Choisir d arbitrairement grand conduit au phénomène largement connu du surapprentissage. Les modèles de dimension supérieure s'adaptent mieux aux données d'apprentissage (en autre $\sigma_d \searrow \sigma$ quand $d \to \infty$), par contre, ils ne sont pas pertinents pour prédire les instances à venir du processus. Observez que, en particulier, d/T (dans le terme de droite de (1.3.6)) peut devenir malencontreusement grand. Plusieurs stratégies pénalisant la dimension d ont été proposées, telles que le critère de l'Erreur de Prédiction Finale (FPE) d'Akaike (1969), le Critère d'Information d'Akaike (AIC) paru dans Akaike (1973), la version avec correction du biais de l'AIC (AICC) de Hurvich and Tsai (1989) et le Critère d'Information Bayésien (BIC) (voir Schwarz (1978) et Akaike (1978)). Nous nous référons à (Brockwell and Davis, 2002, Section 5.5) pour une vue d'ensemble.

Exemple 6 (AR(d)). Le processus autorégressif à valeurs réelles d'ordre d est un cas particulier de l'Exemple 5 (page 9) avec $\theta_j = 0$ pour tout $j > d$, c'est-à-dire

$$X_t = \sum_{j=1}^{d}\theta_j X_{t-j} + \sigma\xi_t , \tag{1.3.7}$$

où $\boldsymbol{\theta} = [\theta_1 \ \dots \ \theta_d]' \in \mathbb{R}^d$. Supposons par ailleurs que la médiane de ξ_t vaut zéro.
Comme dans l'exemple précédent, considérez $Y = (Y_t)_{t\geq 1}$, une copie indépendante de X. Observez que

$$\sigma_j^2 = \inf_{\boldsymbol{\theta}\in\mathbb{R}^j} \mathbb{E}\left[(\boldsymbol{\theta}'\boldsymbol{Y}_{j:1} - Y_{j+1})^2\right] = \sigma^2 \ \text{ pour } \ j \geq d \ .$$

Le meilleur prédicteur du processus étant donné son passé, par rapport à la perte quadratique, est l'espérance conditionnelle $\hat{f}(\boldsymbol{y}_{1:t-1}) = \boldsymbol{\theta}'\boldsymbol{y}_{t-1:t-d}$. De manière générale, pour n'importe quel $\widehat{\boldsymbol{\theta}}$, estimateur $\sigma(X)$-mesurable de $\boldsymbol{\theta}$ nous avons

$$\mathbb{E}\left[\left(\widehat{\boldsymbol{\theta}}'\boldsymbol{Y}_{d:1} - Y_{d+1}\right)^2 \middle| X\right] = \sigma^2 + \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \mathbb{E}\left[\boldsymbol{Y}_{d:1}\boldsymbol{Y}_{d:1}'\right]\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \sigma^2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\Gamma^2 \ , \quad (1.3.8)$$

où $\|\cdot\|_\Gamma$ dénote la norme associée à $\Gamma \in \mathbb{R}^{d\times d}$, la matrice de covariance de Y.
Cela justifie la stratégie de chercher des estimateurs efficaces pour $\boldsymbol{\theta}$ quand on veut prédire Y. Si à la place, on considère la perte ℓ_1, le lien entre estimation et prédiction est moins direct.

$$\mathbb{E}\left[\left|\hat{f}\left(\boldsymbol{Y}_{1:t-1}\middle|\boldsymbol{X}_{1:T}\right) - Y_t\right|\right] = \mathbb{E}\left[\mathbb{E}\left[\left|\hat{f}\left(\boldsymbol{Y}_{1:t-1}\middle|\boldsymbol{X}_{1:T}\right) - \boldsymbol{\theta}'\boldsymbol{Y}_{t-1:t-d} - \sigma\xi_t\right| \middle| \boldsymbol{X}_{1:T}, \boldsymbol{Y}_{1:t-1}\right]\right] \quad (1.3.9)$$

Observez que ξ_t est indépendant de $\boldsymbol{X}_{1:T}, \boldsymbol{Y}_{1:t-1}$. L'espérance conditionnelle dans le terme de droite de l'Équation (2.3.9) est minimisé quand $(\hat{f}_t(\boldsymbol{Y}_{1:t-1}|\boldsymbol{X}_{1:T}) - \boldsymbol{\theta}'\boldsymbol{Y}_{t-1:t-d})/\sigma$ est égal à la valeur médiane de ξ_t. Comme nous supposons que cette valeur médiane vaut zéro, le meilleur prédicteur du processus étant donné son passé, par rapport à la perte ℓ_1, est encore une fois l'espérance conditionnelle $\hat{f}(\boldsymbol{y}_{1:t-1}) = \boldsymbol{\theta}'\boldsymbol{y}_{t-1:t-d}$.
Si les $(\xi_t)_{t\in\mathbb{Z}}$ sont des variables aléatoires gaussiennes centrées standards, le risque de prédiction satisfait

$$\mathbb{E}\left[\left|\hat{f}_t\left(\boldsymbol{Y}_{1:t-1}\middle|\boldsymbol{X}_{1:T}\right) - Y_t\right| \middle| \boldsymbol{X}_{1:T}\right] = \frac{\left(2\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\Gamma\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) + 2\sigma^2\right)^{1/2}}{\pi^{1/2}} \ , \quad (1.3.10)$$

où $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ est un estimateur de $\boldsymbol{\theta}$ qui dépend exclusivement de $\boldsymbol{X}_{1:T}$.

Avoir à disposition une copie du processus à prédire peut s'avérer assez difficile, d'autant plus dans un contexte dépendant. Comme mentionné au début de cette section, la même approche a été utilisée en pratique quand X et Y sont dépendants. Le processus Y peut correspondre, par exemple, à $(X_{T+\Delta+t})_{t\geq 1}$ où Δ est assez grand. Une direction de recherche différente explore la construction de prédicteurs qui ne reposent pas sur des jeux de données indépendants et étudie des bornes rigoureuses pour les risques correspondants.

1.3.2 Sur la prédiction d'un processus dépendant sans ensemble d'apprentissage

Supposez que les observations du processus dépendant $X = (X_t)_{t\in\mathbb{Z}}$ arrivent une après l'autre. Le but de cette section est de présenter des prédicteurs de X_t construits exclusivement à partir de son passé et de fournir des résultats de consistance sous des conditions spécifiques.
L'ensemble d'apprentissage peut être soit vide, soit constitué de toutes les observations $(X_s)_{s\leq 0}$ disponibles avant de commencer la prédiction. Soit $\mathcal{F}$ la filtration naturelle associée à X, c'est-à-dire $\mathcal{F}_t = \sigma(X_s, 1 \leq s \leq t)$. Nous dénotons par $\hat{f}_t$ la fonction qui prédit X_t à partir de $X_1, \ldots, X_{t-1}$ et $\mathcal{H}$, alors, posons $\widehat{X}_t = \hat{f}_t(\boldsymbol{X}_{1:t-1} \mid \mathcal{H}) = \hat{f}_t((X_s)_{s\leq t-1})$. Soit ℓ la perte quadratique.
Si $\mathcal{H}$ est la σ-algèbre triviale, le risque de prédiction et le risque coïncident, étant égaux à

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\hat{f}_t\left(\boldsymbol{X}_{1:t-1}\right) - X_t\right)^2\right]. \tag{1.3.11}$$

Exemple 7 (Processus linéaire variable dans le temps). Supposons que les paramètres $\boldsymbol{\theta}$ et σ varient avec t dans (1.3.7). Une telle généralisation de l'Exemple 6 (page 10) est connue comme processus linéaire avec des coefficients variables dans le temps. Ils sont définis par la représentation

$$X_t = \sum_{j=1}^{\infty}\theta_j(t)\,X_{t-j} + \sigma(t)\,\xi_t\,, \tag{1.3.12}$$

où (ξ_t) est une suite de variables aléatoires centrées avec variance 1.
Comme dans les exemples précédents, l'estimation de $\boldsymbol{\theta}$ donne la clé pour la prédiction. Nous avons besoin de construire $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1 \ldots \widehat{\theta}_d]'$, une fonction de $\mathbb{Z}$ dans $\mathbb{R}^d$, avec $d \in \mathbb{N}^*$. Les méthodes de descente de gradient stochastique (ou en ligne) sont devenues très populaires en raison de leur simplicité intrinsèque et leur efficacité prouvée. L'algorithme primaire, adapté à l'exemple présent, est esquissé dans la suite.

Algorithme 1: Descente de gradient stochastique

paramètres la taille du pas du gradient μ;
initialisation $t = 0, \widehat{\boldsymbol{\theta}}(t) = [0 \ldots 0]'$;
tant que l'entrée X_t *est donnée;*
faire
  $\widehat{\boldsymbol{\theta}}(t+1) = \widehat{\boldsymbol{\theta}}(t) + \mu\left(X_t - \widehat{\boldsymbol{\theta}}'(t)\,\boldsymbol{X}_{t-1:t-d}\right)$;
  retourner $\widehat{\boldsymbol{\theta}}(t+1)$;
  $t = t + 1$;

La convergence de la descente de gradient stochastique a été amplement étudiée dans le cas stationnaire (voir Bottou (1998), et plus récemment Bottou (2012)). Un analyse

pour les suites individuelles est fournie dans Cesa-Bianchi (1999). Par contre, pour la classe de processus décrite dans cet exemple, prouver des résultats en risque de prédiction est difficile ; les résultats disponibles sont plutôt rares.
Comme nous l'avons expliqué dans la Section 1.2.2, les résultats significatifs disponibles pour cette sorte de modèles nécessitent des conditions de régularité spécifiques. Le processus TVAR présenté dans l'Exemple 3 (page 37), où le paramètre $\boldsymbol{\theta}$ est β-Hölder continu avec $\beta \in (0, 1]$ en est un représentant.
Dans ce contexte, supposons que nous avons à disposition assez d'observations $(X_s)_{s\leq 0}$. Le (Moulines et al., 2005, Theorem 2) implique le résultat suivant : soit $\widehat{\boldsymbol{\theta}}$ l'estimation de $\boldsymbol{\theta}$ obtenue à partir de l'algorithme de Moindres Carrés Normalisés (NLMS en anglais, c'est une modification de l'Algorithme 1), il existe une constante $M_1 > 0$ telle que

$$\sup_{1\leq t\leq T} \left(\mathbb{E}\left[\left| \widehat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) \right|^4 \right] \right)^{1/2} \leq M_1 \left(\mu^{1/2} + (T\mu)^{-\beta} \right)^2 .$$

En posant $\hat{f}_t(\boldsymbol{x}_{1:t-1}) = \widehat{\boldsymbol{\theta}}'(t)\, \boldsymbol{x}_{t-1:t-d}$, où $\widehat{X}_t = \widehat{\boldsymbol{\theta}}'(t)\, \boldsymbol{X}_{t-1:t-d}$ nous concluons qu'il existe une constante $M_2 > 0$ telle que

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \left(\hat{f}_t(\boldsymbol{X}_{1:t-1}) - X_t \right)^2 \right] - \frac{1}{T} \sum_{t=1}^{T} \sigma^2\left(\frac{t}{T}\right) \leq M_2 \left(\mu^{1/2} + (T\mu)^{-\beta} \right)^2 . \quad (1.3.13)$$

Ce résultat est valable pour $\beta \in (0, 1]$. Une technique de réduction de biais peut être utilisée afin d'obtenir le même taux de décroissance de (1.3.13) pour $\beta \in (1, 2]$, voir (Moulines et al., 2005, Corollary 9). La taille du pas μ qui minimise le terme de droite de (1.3.13) est proportionnelle à $T^{-2\beta/(2\beta+1)}$. Cette expression contient β qui est usuellement inconnu en pratique.

1.3.3 Optimalité

Dans les sections précédentes nous avons décrit des prédicteurs efficaces qui présentent un risque de prédiction faible avec grande probabilité ou tout simplement un risque faible. Dans la pratique, nous cherchons des bornes supérieures pour ces risques-là, et nous avons besoin qu'elles soient aussi faibles que possible. La définition de "faible" change d'une situation à l'autre. En reprenant l'Exemple 6 (page 6) concernant le processus AR(d), nous avons montré que le risque de prédiction ne pouvait pas être plus petit que σ^2 pour la perte ℓ_2 (voir Équation (1.3.8)) ou plus petit que $(2/\pi)^{1/2}\sigma$ pour la perte ℓ_1 sous l'hypothèse gaussienne (voir Équation (1.3.10)). Nous allons maintenant éclaircir et généraliser les idées relatives à ces bornes inférieures et supérieures. Cette section présente les aspects qui caractérisent l'optimalité d'une procédure de prédiction.
Typiquement, nous cherchons à construire des prédicteurs $\widehat{Z} = (\widehat{Z}_t)_{1\leq t\leq T} \in \mathcal{P}_T$ à partir

d'une collection de prédicteurs indexée par Θ: $\widehat{Z}_\Theta = \{\widehat{Z}_{\boldsymbol{\theta}} = (\widehat{Z}_{\boldsymbol{\theta},t})_{1\le t\le T} \in \mathcal{P}_T, \boldsymbol{\theta} \in \Theta\}$. Cela ne veut pas pour autant dire que $\widehat{Z} \in \widehat{Z}_\Theta$ car $\widehat{Z}_\Theta$ peut être strictement contenu dans $\mathcal{P}_T$. Soit $\widehat{Z}_* = (\widehat{Z}_{*,t})_{1\le t\le T}$ tel que

$$\widehat{Z}_* \in \underset{\widehat{Z}\in\mathcal{P}_T}{\arg\inf} R_T\left(\widehat{Z}\right) ,$$

où l'inf est pris sur tous les prédicteurs possibles de $(Z_t)_{1\le t\le T}$ (voir la Définition 4). Si $\widehat{Z}_*$ correspond à $\widehat{Z}_{\boldsymbol{\theta}}$ avec $\boldsymbol{\theta} \in \Theta$, nous dirons que le modèle est bien spécifie. Autrement il est dit mal spécifié.

1

Depuis notre point de vue, la comparaison avec le meilleur des prédicteurs est plus informative que le risque en lui même or elle mesure à quel point nous pouvons bien prédire tout en exploitant la connaissance sur le processus dont nous disposons. La décomposition suivante du risque soulève le souci récurrent du compromis biais-variance

$$R_T\left(\widehat{Z}\right) - R_T\left(\widehat{Z}_*\right) = \underbrace{\left(R_T\left(\widehat{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right)\right)}_{\text{regret}} + \underbrace{\left(\inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right) - R_T\left(\widehat{Z}_*\right)\right)}_{\text{erreur d'approximation}} . \tag{1.3.14}$$

Le premier terme entre parenthèses dans (1.3.14) correspond à ce que nous appelons le regret (ou regret lié au meilleur prédicteur). Il mesure la pertinence de notre choix dans Z_Θ. Le deuxième terme désigne l'erreur d'approximation et évalue la pertinence de la classe Z_Θ. D'un côté, à des classes Z_Θ plus larges correspondent des erreurs d'approximation plus faibles mais des regrets plus importants. D'un autre côté, des valeurs plus larges de T n'ont pas d'impact dans l'erreur d'approximation mais d'habitude entrainent des regrets plus faibles. Une question cruciale est celle du compromis entre la taille de Z_Θ et T. D'un point de vue pratique, un troisième terme d'erreur peut apparaitre, il est inhérent à la méthode numérique qu'on choisi afin de calculer $\widehat{Z}$. Nous pouvons réécrire le regret en incluant $\widetilde{Z}$, l'approximation numérique de $\widehat{Z}$ que nous venons de mentionner

$$\underbrace{R_T\left(\widetilde{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right)}_{\text{regret numérique}} = \underbrace{\left(R_T\left(\widetilde{Z}\right) - R_T\left(\widehat{Z}\right)\right)}_{\text{erreur d'approximation numérique}} + \underbrace{\left(R_T\left(\widehat{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right)\right)}_{\text{regret}} . \tag{1.3.15}$$

Dans les exemples 5, 6 et 7 (pages 9, 10 et 12 respectivement), le fait d'avoir des modèles bien spécifiés équivaut à supposer que le paramètre (ou la fonction) $\boldsymbol{\theta}$ qui génère le processus se trouve dans Θ. Nous avons $R_T(\widehat{Z}_*) = \sigma^2$ dans les exemples 5 et 7 et les bornes respectives (1.3.6) et (1.3.13) sont déjà exprimées en forme de regret. Dans l'Exemple 6, $R_T(\widehat{Z}_*) = (2/\pi)^{1/2}\sigma$.

Dans cette contribution, le regret est notre mesure de qualité d'un prédicteur. Les deux sections suivantes introduisent les inégalités oracle et l'approche minimax, toutes les deux concernant le regret. Ensuite, nous présentons certains outils du MCMC liés à l'erreur d'approximation numérique.

1.3.3.1 Inégalités oracle

Les notions d'*oracle* et *inégalités oracle* furent introduites par Donoho and Johnstone (1998). Les inégalités oracle fournissent des bornes supérieures pour le regret d'une procédure statistique en fonction de T (voir (Tsybakov, 2009, Section 1.8)). Ces bornes sont aussi, de manière générale, dépendantes des paramètres définissant la classe de modèles imposé sur Z. Dorénavant, $\mathcal{M}_\lambda$ dénote la classe de processus à laquelle Z appartient. Elle est indexée par l'hyperparamètre $\boldsymbol{\lambda}$. Dans l'Exemple 5 (page 9) nous pouvons considérer que $\mathcal{M}_\lambda$ est la collection de tous les processus qui satisfont l'Équation (1.3.5) où les paramètre qui les génèrent se trouvent dans $s_\infty(\delta) \times \{\sigma\} = \{\{\boldsymbol{\theta} \in \mathbb{R}^{\mathbb{N}^*}, 1 - \sum_{j=1}^{\infty} \theta_j z^j \neq 0, \text{pour } |z| \leq \delta^{-1}\} \cap \{|\boldsymbol{\theta}|_1 < \infty\}\} \times \{\sigma\}$, et $\boldsymbol{\lambda} = (\delta, \sigma) \in \mathbb{R}_+^{*2}$.
Nous concentrons particulièrement notre attention sur des inégalités oracle valables uniformément sur $\mathcal{M}_\lambda$, c'est-à-dire, quand il existe $M_\lambda > 0$ uniquement dépendant de $\boldsymbol{\lambda}$ et une suite $(\psi_{T,\lambda})_{T\geq 1}$ tels que, pour tout $T \geq 1$

$$\sup_{Z \in \mathcal{M}_\lambda} \left\{ R_T\left(\widehat{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right) \right\} \leq M_\lambda \psi_{T,\lambda} \,, \tag{1.3.16}$$

et aussi quand l'inégalité précédente est valable avec grande probabilité, cela signifie qu'il existe $M_\lambda > 0$ qui dépend uniquement de $\boldsymbol{\lambda}$ et $(\psi_{T,\lambda,\varepsilon})_{T\geq 1}$ tel que pour tout $\varepsilon \in (0,1)$ et $T \geq 1$, avec probabilité au mois $1 - \varepsilon$ nous avons

$$\sup_{Z \in \mathcal{M}_\lambda} \left\{ R_T\left(\widehat{Z}\middle|\mathcal{H}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\middle|\mathcal{H}\right) \right\} \leq M_\lambda \psi_{T,\lambda,\varepsilon} \,.$$

L'ensemble Θ peut être arbitraire. Cependant, ils existent des choix convenables conditionnés par l'information sur $\boldsymbol{\lambda}$ dont on dispose, comme par exemple l'ensemble de toutes les vecteurs de $\mathbb{R}^d$ qui génèrent des processus autorégressifs non-explosifs dans l'Exemple 6 (page 10) ou une classe de fonctions Hölder stables dans l'Exemple 7 (page 10). Le prédicteur $\widehat{Z}_{\boldsymbol{\theta}^*}$, où $\boldsymbol{\theta}^* = \arg\inf_{\boldsymbol{\theta}\in\Theta} R_T(\widehat{Z}_{\boldsymbol{\theta}})$, est appelé projection oracle, car il donne la meilleure prédiction de Z dans Θ.

Hyperparamètre des processus TVAR :

Le cas des processus TVAR, décrit dans l'Exemple 3 (page 6), revêt un intérêt tout particulier. Rappelons que les résultats statistiques existants (y compris l'inégalité oracle (1.3.13)) exploitent la régularité de $\boldsymbol{\theta}$ (donnée par β et R) et reposent aussi sur cette stabilité (à travers de β, R et δ). Les paramètres décrivant ces aspects, aussi comme la borne de σ, définissent $\boldsymbol{\lambda}$.
Soient $\beta > 0, \delta \in (0,1), R > 0, \rho \in (0,1]$ et $\sigma_+ > 0$. Nous disons que X appartient à $\mathcal{M}_\lambda$ avec

$$\boldsymbol{\lambda} = (\beta, R, \delta, \rho, \sigma_+) \,, \tag{1.3.17}$$

si X est un processus TVAR généré par une fonction $\boldsymbol{\theta} \in s_d(\delta) \cap \Lambda_d(\beta, R)$ (voir les équations (2.2.8) et (2.2.9)), avec $\sigma \in [\rho\sigma_+, \sigma_+]$.
L'inégalité (1.3.13) qui correspond à l'Exemple 7 est valable uniformément pour tous les processus $X \in \mathcal{M}_\lambda$. Elle remplit la définition d'inégalité oracle donnée par (1.3.16).

1.3.3.2 Minimax et adaptabilité

Une première et plus courante question est à quelle vitesse notre méthode approxime la projection oracle. La réponse est apportée par des inégalités oracle comme expliqué dans la Section 1.3.3.1. Il y a une autre question qui complète la première : à quelle vitesse la prédiction peut être faite dans $\mathcal{P}_T$? La réponse arrive de la main du regret minimax.
Le regret de prédiction minimax est défini selon

$$\inf_{\widehat{Z}\in\mathcal{P}_T}\sup_{Z\in\mathcal{M}_\lambda}\left\{R_T\left(\widehat{Z}\right)-\inf_{\theta\in\Theta}R_T\left(\widehat{Z}_\theta\right)\right\}\ . \tag{1.3.18}$$

Aucun prédicteur ne peut mimer la projection oracle plus vite que n'importe quelle borne inférieure de (1.3.18). On observe que si le modèle est bien spécifié, l'expression (1.3.18) se transforme en

$$\inf_{\widehat{Z}\in\mathcal{P}_T}\sup_{Z\in\mathcal{M}_\lambda}\left\{R_T\left(\widehat{Z}\right)-R_T\left(\widehat{Z}_*\right)\right\}\ .$$

Dans ce contexte, toutes les bornes inférieures sont non négatives.
Nous supposons qu'il existe une constante $m_\lambda > 0$ qui dépend seulement de λ et une suite $(\psi_{T,\lambda})_{T\geq 1}$ telle que

$$\inf_{\widehat{Z}\in\mathcal{P}_T}\sup_{Z\in\mathcal{M}_\lambda}\left\{R_T\left(\widehat{Z}\right)-\inf_{\theta\in\Theta}R_T\left(\widehat{Z}_\theta\right)\right\}\geq m_\lambda\psi_{T,\lambda}\ . \tag{1.3.19}$$

Par conséquent, le plus vite que nous pouvons prédire est aussi borné inférieurement par la suite $(\psi_{T,\lambda})_{T\geq 1}$ dans (1.3.19). Un prédicteur $\widehat{Z}$ tel que l'inégalité (1.3.16) a lieu pour la même suite $(\psi_{T,\lambda})_{T\geq 1}$ est dit optimal ou minimax en vitesse.
Il existe plusieurs études portant sur des problèmes minimax dans des contextes différents : estimation non paramétrique (voir Gill and Levit (1995) et Nemirovskiĭ (1990)), estimation de densité (voir Čencov (1962) et Yang and Barron (1999)), estimation de densité dans un point fixe (voir Farrell (1972)). De la même manière, nous recommandons Birgé (1983) et (Tsybakov, 2009, Chapter 2) et les références qui y sont. Les techniques développées auparavant offrent une approche pour résoudre le problème de prédiction minimax.
La construction de prédicteurs peut reposer sur la connaissance du paramètre λ qui définit la classe $\mathcal{M}_\lambda$, qu'en pratique nous ignorons. Dans l'Exemple 7 (page 12), afin d'obtenir un prédicteur avec le meilleur taux de convergence nous devons, a priori, connaitre β (voir Équation (1.3.13) et la remarque qui suit).
De façon plus générale, l'arsenal de prédicteurs minimax en vitesse ou optimaux construit quand nous connaissons λ peut s'avérer inutile si cette information n'est pas disponible. Le comportement d'un prédicteur $\widehat{Z}$ peut varier d'une classe $\mathcal{M}_\lambda$ à une autre. Les

méthodes qui contournent ce souci, et qui sont minimax en vitesse pour tout λ dans un certain ensemble, son appelées adaptatives.
Soit Λ l'ensemble des valeurs possibles de λ. Nous disons que le prédicteur $\widehat{Z}$ est Λ minimax adaptatif si pour tout $\lambda \in \Lambda$ il est minimax en vitesse. Dit d'une autre manière, pour tout $\lambda \in \Lambda$, il existe $M_\lambda > 0$ dépendant exclusivement de λ tel que pour tout $T \geq 1$

$$\sup_{Z \in \mathcal{M}_\lambda} \left\{ R_T\left(\widehat{Z}\right) - \inf_{\theta \in \Theta} R_T\left(\widehat{Z}_\theta\right) \right\} \leq M_\lambda \psi_{T,\lambda} \ ,$$

où la suite $(\psi_{T,\lambda})_{T \geq 1}$ satisfait (1.3.19).
Les méthodes adaptatives sont convenables car elles convergent à la même vitesse du meilleur des prédicteurs, sans pour autant nécessiter d'information précise sur le processus à prédire.
Les méthodes minimax adaptatives datent des années 1980s. Depuis, des nombreux travaux étudiant des problèmes différents sont apparus. Nous nous référons par exemple à Efroimovich and Pinsker (1984); Lepskiĭ (1991); Barron and Cover (1991); Donoho and Johnstone (1998); Birgé and Massart (2000); Yang (2000a).

1.3.3.3 Méthodes Monte Carlo par chaines de Markov

Dans la pratique, le calcul numérique du prédicteur $\widehat{Z}$ peut conduire à un comportement non expliqué par la borne (1.3.14). Si on pouvait calculer exactement $\widehat{Z}$ la question ne se poserait pas ; autrement il est utile d'étudier comment $\widetilde{Z}$, l'approximation numérique de $\widehat{Z}$, mime la projection oracle (voir Équation (1.3.15)).
Le prédictor $\widehat{Z}$ est quelques fois donné par une intégrale (voir Dalalyan and Tsybakov (2012)). Les amplement utilisés méthodes Monte Carlo par chaines de Markov (MCMC) proposent une boite à outils pour l'approximer (voir Cappé et al. (2005) et Meyn and Tweedie (2009)). Pourtant, il est crucial de borner le nombre d'itérations que l'algorithme nécessite afin d'atteindre une précision numérique du même ordre que le risque de prédiction. Un article de Łatuszyński and Niemiro (2011) contient plusieurs résultats qui évaluent la précision d'une approximation MCMC en fonction du nombre d'itérations.
Supposez qu'il existe une fonction g telle que

$$\widehat{Z}_t = \int g(\boldsymbol{u}) \, \pi_0(\mathrm{d}\boldsymbol{u}) \ , \tag{1.3.20}$$

où $\boldsymbol{u}$ appartient à un certain espace $\mathcal{U}$ (nous supposons qu'il est un sous-espace de $\mathbb{R}^d$ avec $d > 0$) équipé de la mesure π_0. Nous considérons une chaine de Markov $U = (U_i)_{i \geq 0}$ avec distribution invariante π_0. Dénotons par μ la distribution de probabilité de U. Nous approximons l'intégrale (1.3.20) par

$$\widehat{Z}_{t,n} = \frac{1}{n} \sum_{i=0}^{n-1} g(U_i) \tag{1.3.21}$$

D'un côté, le comportement asymptotique de $\widehat{Z}_{t,n}$ est souvent étudié à travers le Théorème Central Limite (TCL) pour des chaines de Markov (voir Geyer (1992), Jones (2004) et Roberts and Rosenthal (2004)). Des intervalles de confiance sont établis grâce au TCL (nous nous référons à Geyer (1992); Flegal and Jones (2010); Jones and Hobert (2001)). De l'autre côté, Łatuszyński and Niemiro (2011) propose une borne inférieure explicite pour n qui garantit ce qui suit

$$\mu\left(\left|\widehat{Z}_{t,n} - \widehat{Z}_t\right| \leq \alpha\right) \geq 1 - \varepsilon \,,$$

pour $\alpha, \varepsilon > 0$, où $\mu(A)$ représente la probabilité de A par rapport à la distribution μ.
Cette borne inférieure dépend de α, ε, la fonction g et d'une certaine condition de dérivé supposée pour la chaine de Markov U (voir (Łatuszyński and Niemiro, 2011, Theorem 3.1)). La condition de dérivé implique (sous des conditions adéquates) l'ergodicité géométrique (voir Meyn and Tweedie (2009) et (Baxendale, 2005, Theorem 1.1)). Il s'agit de l'ingrédient principal que la preuve du (Łatuszyński and Niemiro, 2011, Theorem 3.1) nécessite.
Les chaines de Markov U, les plus convenables, sont celles qui convergent plus rapidement à la distribution invariante π_0 et qui exhibent les bornes inférieures les plus petites pour n. Dit d'une autre manière, elles permettent d'approximer $\widehat{Z}_t$ à un niveau α (et particulièrement pour $\alpha \propto \psi_{\lambda,T}$ tel comme défini dans (1.3.16)) en moins d'itérations.

1.4 AGRÉGATION

Après avoir introduit les modèles (Section 1.2) et avoir explicité ce que nous cherchons avec une procédure de prédiction (Section 1.3), nous présentons ici une approche que nous utilisons afin de proposer nos prédicteurs.
Une des machineries générales pour aborder les problèmes de prédiction sont les méthodes d'agrégation. Elles sont étudiées depuis 25 ans. Les techniques d'agrégation demeurent dans le carrefour de deux communautés faisant de l'apprentissage statistique, voir Vovk (1990); Littlestone and Warmuth (1994); Haussler et al. (1998) pour la première et les travaux séminaux de Barron (1987); Catoni (1997, 2004); Juditsky and Nemirovski (2000); Yang (2000a, 2004); Leung and Barron (2006) pour la deuxième. Voir (Giraud, 2015, Chapter 3) pour un aperçu récent.
Des algorithmes populaires d'agrégation tels que le Boosting (Freund (1995)), le Bagging (Breiman (1996)) et les Forets Aléatoires (Amit and Geman (1997)) ont été amplement appliqués en pratique avec du succès.
Soit Θ un ensemble d'indexes, possiblement non dénombrable, équipé avec une σ-algèbre (à définir) et soit π une mesure définie sur ce couple-là appelée *prior*. Supposez que les observations appartiennent à $\mathcal{X} \subseteq \mathbb{R}$. Supposez additionnellement que $\pi(\Theta) = \int_\Theta \pi(\mathrm{d}\boldsymbol{\theta}) < \infty$. Il nous est donnée une collection $\{(\widehat{x}_t^{(\boldsymbol{\theta})})_{1 \leq t \leq T}, \boldsymbol{\theta} \in \Theta\}$, que nous appelons prédicteurs (il s'agit d'une terminologie, ils ne sont pas nécessairement des prédicteurs dans le sens de la Définition 4). Notre premier objectif est d'obtenir un nouveau prédicteur qui

prédise aussi précisément ou plus précisément que la meilleure combinaison convexe de $\{(\widehat{x}_t^{(\boldsymbol{\theta})})_{1\le t\le T}, \boldsymbol{\theta}\in\Theta\}$ sans pour autant connaitre laquelle est elle. Un objectif moins ambitieux est de concevoir un prédicteur qui se comporte de façon similaire ou mieux que le meilleur de la collection qui nous en est fournie.
Définissons le simplex

$$\mathcal{S}_\Theta = \left\{ \boldsymbol{s} = (s_{\boldsymbol{\theta}}, \boldsymbol{\theta}\in\Theta) \in \mathbb{R}_+^\Theta : \int_\Theta s_{\boldsymbol{\theta}} \pi(\mathrm{d}\boldsymbol{\theta}) = 1 \right\}. \tag{1.4.1}$$

Soit $\ell : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ une fonction perte. Nous construisons le nouveau prédicteur en utilisant une collection $\{(\alpha_{\boldsymbol{\theta},t})_{1\le t\le T}, \boldsymbol{\theta}\in\Theta\}$ telle que pour tout $1\le t\le T$, $\boldsymbol{\alpha}_t = (\alpha_{\boldsymbol{\theta},t}, \boldsymbol{\theta}\in\Theta) \in \mathcal{S}_\Theta$. Nous présentons formellement nos deux buts.
Le premier est de construire $\{(\alpha_{\boldsymbol{\theta},t})_{1\le t\le T}, \boldsymbol{\theta}\in\Theta\}$ tel que

$$\frac{1}{T}\sum_{t=1}^T \ell\left(\int_\Theta \alpha_{\boldsymbol{\theta},t}\widehat{x}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta}), x_t\right) - \inf_{\nu\in\mathcal{S}_\Theta}\frac{1}{T}\sum_{t=1}^T \ell\left(\int_\Theta \nu_{\boldsymbol{\theta}}\widehat{x}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta}), x_t\right), \tag{1.4.2}$$

est aussi petit que possible (connu comme le problème de la borne du regret convexe).
Le second objectif est de proposer $\{(\alpha_{\boldsymbol{\theta},t})_{1\le t\le T}, \boldsymbol{\theta}\in\Theta\}$ tel que

$$\frac{1}{T}\sum_{t=1}^T \ell\left(\int_\Theta \alpha_{\boldsymbol{\theta},t}\widehat{x}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta}), x_t\right) - \inf_{\boldsymbol{\theta}\in\Theta}\frac{1}{T}\sum_{t=1}^T \ell\left(\widehat{x}_t^{(\boldsymbol{\theta})}, x_t\right), \tag{1.4.3}$$

est aussi petit que possible (connu comme le problème de la borne du regret du meilleur prédicteur).
Les expressions (1.4.2) et (1.4.3) rappellent le regret introduit dans (1.3.14). Leurs versions stochastiques sont détaillées dans la Section 1.4.2. Quand ℓ est la perte quadratique, deux stratégies d'agrégation par poids exponentiels sont extensivement étudiées.

Stratégie 1 : construisant des poids à partir du gradient de la perte quadratique :

La première stratégie consiste à définir pour tout $\boldsymbol{\theta}\in\Theta$ et $t = 1,\dots,T$, les poids $\widehat{\alpha}_{\boldsymbol{\theta},t}$ selon

$$\widehat{\alpha}_{\boldsymbol{\theta},t} = \frac{\exp\left(-2\eta\sum_{s=1}^{t-1}\left(\int_\Theta \widehat{\alpha}_{\boldsymbol{\theta}_1,s}\widehat{x}_s^{(\boldsymbol{\theta}_1)}\pi(\mathrm{d}\boldsymbol{\theta}_1) - x_s\right)\widehat{x}_s^{(\boldsymbol{\theta})}\right)}{\int_\Theta \exp\left(-2\eta\sum_{s=1}^{t-1}\left(\int_\Theta \widehat{\alpha}_{\boldsymbol{\theta}_1,s}\widehat{x}_s^{(\boldsymbol{\theta}_1)}\pi(\mathrm{d}\boldsymbol{\theta}_1) - x_s\right)\widehat{x}_s^{(\boldsymbol{\theta}_2)}\right)\pi(\mathrm{d}\boldsymbol{\theta}_2)}, \tag{1.4.4}$$

avec la convention que la somme de zéro éléments est nulle, donc $\widehat{\alpha}_{\boldsymbol{\theta},1} = 1/\pi(\Theta)$ pour tout $\boldsymbol{\theta}$.
Le paramètre $\eta > 0$, est usuellement appelé *taux d'apprentissage*. Nous fixons sa valeur en fonction de la nature des observations. Cette stratégie assure des garanties pour le

prédicteur agrégé comparé avec la meilleure des combinaisons convexes constantes de prédicteurs. Son inconvénient est que le regret est de l'ordre de $T^{-1/2}$ (les détails peuvent être trouvés dans la preuve de l'inégalité (4.2.12) du Lemme 5, page 91).

Stratégie 2 : construisant des poids à partir de la perte quadratique : La deuxième strategie consiste à définir pour tout $\boldsymbol{\theta} \in \Theta$ et $t = 1, \ldots, T$, les poids $\widehat{\alpha}_{\boldsymbol{\theta},t}$ selon

$$\widehat{\alpha}_{\boldsymbol{\theta},t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(\boldsymbol{\theta})} - x_s\right)^2\right)}{\int_{\Theta} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(\boldsymbol{\theta}_1)} - x_s\right)^2\right) \pi(\mathrm{d}\boldsymbol{\theta}_1)}, \tag{1.4.5}$$

avec, à nouveau, la convention que la somme de zéro éléments est nulle. Quand les observations et les prédicteurs sont bornés, par exemple, cette stratégie profite d'un regret qui décroit comme T^{-1} pour un η bien choisi (voir l'inégalité (4.2.13) dans Lemme 5, page 91). Le résultat provient de l'exp-concavité de la perte quadratique (nous nous référons à (Cesa-Bianchi and Lugosi, 2006, Section 3.3) et (Catoni, 2004, Proposition 2.2.1)). Elle est étroitement liée à des nombreux développements dans le cadre stochastique ; voir Exemple 10. Le regret dans ce cas-là est calculé par rapport au meilleur prédicteur. Depuis cette perspective, le résultat est plus faible que celui obtenu en utilisant les poids (1.4.4).

1.4.1 Prédiction séquentielle

En contrastant avec le point de vue statistique, la théorie de suites individuelles ne suppose pas que les observations sont la réalisation d'un processus stochastique (voir Cesa-Bianchi and Lugosi (2006)). Depuis cette perspective, l'apprentissage en ligne (aussi appelé prédiction séquentielle) propose une boite à outils d'algorithmes pour aborder des problèmes en apprentissage statistique. Nous nous référons aux travaux de Foster (1991); Auer et al. (2002); Vovk (2006); Stoltz (2011); Gerchinovitz (2013) en régression en ligne pour des séquences arbitraires.

Exemple 8. Supposez que $\Theta = \{1, \ldots, N\}$ et que $\pi(k) = 1$ pour tout $k = 1, \ldots, N$. L'Équation (1.4.5) se transforme en

$$\widehat{\alpha}_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(i)} - x_s\right)^2\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(k)} - x_s\right)^2\right)}. \tag{1.4.6}$$

Admettez que les observations et les prédictions appartiennent à l'intervalle $[-B, B]$ avec $B > 0$. Dans ce contexte, (Cesa-Bianchi and Lugosi, 2006, Theorem 3.2 et

Proposition 3.1) assurent que pour tout $0 < \eta < 1/(8B^2)$ et $T > 0$

$$\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\widehat{\alpha}_{i,t}\widehat{x}_t^{(i)} - x_t\right)^2 - \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{x}_t^{(i)} - x_t\right)^2 \le \frac{\log N}{T\eta}\,, \tag{1.4.7}$$

où $\widehat{\alpha}_{i,t}$ est défini par l'Équation (2.4.6). Nous nous référons à l'inégalité (4.2.13) du Lemme 5 pour une généralisation.

Les contextes de prédiction séquentielle et stochastique ne partagent seulement que des techniques. Le premier permet aussi de comprendre quels termes des garanties obtenues pour le deuxième (inégalités oracle) s'expliquent par les suppositions statistiques et quels termes sont inhérentes à la procédure (l'agrégation dans notre cas).

1.4.2 Prédiction stochastique

Comme expliqué dans la Section 1.3.3.1, lorsque on impose un modèle stochastique sur les observations, nous posons le problème de la borne du regret en termes d'espérance et d'espérance conditionnelle. Dans cette section nous présentons des résultats intéressants dans ce cadre.
Étant donné $\{\widehat{X}^{(\boldsymbol{\theta})}, \boldsymbol{\theta} \in \Theta\}$, avec $\widehat{X}^{\boldsymbol{\theta}} = (\widehat{X}_t^{(\boldsymbol{\theta})})_{1\le t\le T}$, et $\boldsymbol{\alpha} = \{(\alpha_{\boldsymbol{\theta},t})_{1\le t\le T}, \boldsymbol{\theta} \in \Theta\}$, avec $\boldsymbol{\alpha}_t = (\alpha_{\boldsymbol{\theta},t}, \boldsymbol{\theta} \in \Theta) \in \mathcal{S}_\Theta$, nous dénotons $\widehat{X}^{[\alpha]} = (\widehat{X}_t^{[\alpha_t]})_{1\le t\le T}$ le prédicteur agrégé défini selon

$$\widehat{X}_t^{[\alpha_t]} = \int_\Theta \alpha_{\boldsymbol{\theta},t}\widehat{X}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta})\,. \tag{1.4.8}$$

Pour un $\boldsymbol{\nu} \in \mathcal{S}_\Theta$ nous utilisons la même notation $\widehat{X}^{[\nu]} = (\widehat{X}_t^{[\nu]})_{1\le t\le T}$, où

$$\widehat{X}_t^{[\nu]} = \int_\Theta \nu_{\boldsymbol{\theta}}\widehat{X}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta})\,. \tag{1.4.9}$$

Observez que, contrastant avec (1.4.8), dans l'expression (1.4.9) le poids $\nu_{\boldsymbol{\theta}}$ est le même pour tout $t = 1, \ldots, T$.
Le problème de la borne du regret convexe cherche à borner supérieurement

$$R_T\left(\widehat{X}^{[\alpha]}\right) - \inf_{\nu\in\mathcal{S}_\Theta} R_T\left(\widehat{X}^{[\nu]}\right)\,.$$

Dans le cas de la borne du regret du meilleur prédicteur, l'expression à borner supérieurement est

$$R_T\left(\widehat{X}^{[\alpha]}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{X}^{(\boldsymbol{\theta})}\right)\,.$$

Pour un souci de breveté, nous n'écrivons pas les bornes du regret conditionnel qui y correspondent.

Les poids $\alpha_{\theta,t}\pi(\mathrm{d}\theta)$ peuvent dépendre de $(X_s)_{s<t}$, $\{\widehat{X}_s^{(\theta)}, s < t, \theta \in \Theta\}$ et aussi de la σ-algèbre d'apprentissage $\mathcal{H}$ (possiblement dépendante des observations). Cette aléa extra apparue dans les poids a un gout PAC-Bayésien.
La théorie de l'apprentissage Probablement Approximativement Correct (PAC), introduite par Valiant (1984), fourni des garanties dans l'erreur d'approximation d'une statistique qui sont satisfaites avec forte probabilité par rapport à la représentativité de l'ensemble d'apprentissage (dans notre contexte, nous pouvons l'interpréter comme $\mathcal{H}$). La modélisation statistique Bayésienne repose sur la distribution prior que nous imposons aux paramètres inconnus. Les inégalités PAC-Bayésiennes s'inspirent de ces deux théories et furent introduites par McAllester (1999). Quelques années plus tard, Audibert (2004), Catoni (2004) et Dalalyan and Tsybakov (2008) ont montré des inégalités PAC-Bayésiennes pour des procédures d'agrégation.

Exemple 9. Considérez le cadre classique où un échantillon i.i.d. $((X_i, Y_i))_{1\le i\le n}$ nous est donné et que nous souhaitons prédire le Y_{n+1} en fonction de X_{n+1} et de l'ensemble d'apprentissage. Supposons que nous pouvons reposer sur un ensemble de prédicteurs $\widehat{Y}^{(k)} : \mathcal{X} \to \mathcal{Y}$ indexés avec $k \in \Theta = \{1, \dots, N\}$, et que nous mettons $\pi(k) = 1$ pour tout k. Supposons en plus qu'il existe $B > 0$ qui borne presque surement Y et les prédictions $\widehat{Y}^{(k)}$ pour tout k.
Il existe $M > 0$ tel que le prédicteur $\widehat{Y}^{[\alpha]}$ construit avec les poids exponentiels $\boldsymbol{\alpha}$ calculé dans (Audibert, 2004, Section 4.2.2, Chapter 1) satisfait pour tout $\varepsilon > 0$

$$R\left(\widehat{Y}^{[\alpha]}\middle|\mathcal{H}\right) - \inf_{\nu\in\mathcal{S}_\Theta} R\left(\widehat{X}^{[\nu]}\middle|\mathcal{H}\right) \le M\left(\frac{\log\left(N\log\left(2n\right)/\varepsilon\right)}{n}\right)^{1/2} + M\left(\frac{\log\left(N\log\left(2n\right)/\varepsilon\right)}{n}\right). \quad (1.4.10)$$

Exemple 10. Supposez maintenant qu'on observe $((X_t, Y_t))_{t\ge1}$, instances d'un processus possiblement dépendant et non-stationnaire. Le contexte est similaire à celui décrit dans Section 1.3.2. Considérez la décomposition suivante

$$Y_t = \mathbb{E}\left[Y_t\middle|X_t, (X_s, Y_s)_{s\le t-1}\right] + \xi_t\,. \quad (1.4.11)$$

Admettez qu'il nous est donnée une collection dénombrable de prédicteurs $\widehat{Y}^{(k)} : \mathcal{X} \to \mathcal{Y}$ indexée par $k \in \Theta = \mathbb{N}^*$ et à une distance bornée de l'espérance conditionnelle (voir le terme de droite de l'Équation (1.4.11)). Sous une condition de moment exponentiel sur le bruit ξ, (Yang, 2004, Theorem 5) assure que pour $\eta > 0$ assez petit, le prédicteur $\widehat{Y}^{[\widehat{\alpha}]}$, construit avec les poids $\widehat{\alpha}$ défini comme dans (1.4.5), satisfait

$$R\left(\widehat{Y}^{[\widehat{\alpha}]}\right) \le \inf_{k\ge1}\left\{\frac{\log\left(1/\pi_k\right)}{\eta T} + R\left(\widehat{Y}^{(k)}\right)\right\},$$

où $\sum_{k=1}^{\infty} \pi_k = 1$.

L'optimalité du terme résiduel de l'agrégation (par exemple, le terme de droite de (1.4.10)) a été étudié, particulièrement dans le cadre i.i.d. Nous présentons une borne inférieure du terme résiduel bien connue dans le contexte de l'estimation d'une fonction de régression.

Exemple 11. Considérez le modèle de régression

$$Y_i = f(X_i) + \xi_i \,,$$

où les $(X_i)_{1\leq i\leq n}$ sont des vecteurs aléatoires i.i.d. et les $(\xi_i)_{1\leq i\leq n}$ sont des Gaussiennes réelles, i.i.d., centrées, et indépendantes de $(X_i)_{1\leq i\leq n}$. Soit $\mathcal{F}_0 = \{f : |f|_\infty \leq L\}$, avec $L > 0$ et $\Theta = \{1, \dots, N\}$. Sous des suppositions assez souples (Tsybakov, 2003, Theorem 2) garantit qu'il existe $c > 0$ tel que

$$\sup_{f_1,\dots,f_N\in\mathcal{F}_0} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}_0} \left\{ \mathbb{E}\left[\left(\hat{f}_n(X) - f(X)\right)^2\right] - \min_{\theta\in\mathcal{S}_\Theta} E\left[\left(\sum_{k=1}^{N} \theta_k f_k(X) - f(X)\right)^2\right] \right\} \geq c\psi_n \,,$$

où l'inf est pris sur tous les estimateurs, c'est-à-dire les fonctions réelles $\sigma((X_i, Y_i)_{1\leq i\leq n})$-mesurables $\hat{f}_n$, X est supposé être indépendant de $\sigma((X_i, Y_i)_{1\leq i\leq n})$ et

$$\psi_n = \begin{cases} N/n & , \text{if } N \leq n^{1/2} \,, \\ \left((1/n)\log\left(1 + N/n^{1/2}\right)\right)^{1/2} & , \text{if } N > n^{1/2} \,. \end{cases}$$

Afin d'approfondir dans ce sujet, nous nous référons aussi aux contributions de Juditsky and Nemirovski (2000); Yang (2004); Audibert (2009).

1.5 QUESTIONS DE LA THÈSE

Le Chapitre 3 porte sur les décalages de Bernoulli causales. Notre question principal est comment prédire un CBS $Y = (Y_t)_{t\geq 1}$ étant donné un ensemble d'apprentissage $X = (X_t)_{1\leq t\leq T}$ et une collection de prédicteurs possiblement infinie $\{f_\theta, \theta \in \Theta\}$ tel comme présenté dans les sections 1.3.1.1 et 1.3.3. En bref, nous présentons une inégalité oracle PAC-Bayésienne du risque de prédiction et une inégalité oracle PAC-Bayésienne qui s'applique à l'approximation numérique du prédicteur agrégé.
Dans le Chapitre 4 nous introduisons les processus sous-linéaires, qui sont, en général, dépendants, non-stationnaires et non uniformément bornés. En considérant un nombre fini de prédicteurs, et sans ensemble d'apprentissage (comme dans la Section 1.3.2) nous étudions comment prédire cette sorte de processus en nous utilisant l'agrégation.

Dans le cas particulier des processus auto-régressifs variables dans le temps (TVAR) localement stationnaires (voir Exemple 3, page 6) nous cherchons des prédicteurs minimax adaptatifs. Soient $R, \delta, \rho, \sigma_+ > 0$, et soit J un sous-ensemble compact de $\mathbb{R}_+^*$. Plus précisément, nous ciblons des prédicteurs Λ minimax adaptatifs (voir Section 1.3.3.2) où $\Lambda = \{(\beta, R, \delta, \rho, \sigma_+) : \beta \in J\}$.
Comme observé dans l'Exemple 7 (page 12), des prédicteurs minimax en vitesse sont disponibles pour des processus TVAR quand la régularité β appartient à $(0, 2]$. Le Chapitre 5 propose en particulier des nouveaux prédicteurs minimax en vitesse quand $\beta \geq 2$. Le but du chapitre est plus général, nous y étudions le problème de la régression dans un contexte localement stationnaire.

1

1.6 RÉSULTATS PRINCIPAUX

1.6.1 Décalages de Bernoulli Causales

Dans le Chapitre 3 nous supposons avoir observé un CBS $(X_t)_{1\leq t\leq T}$ distribué selon P dont une copie indépendante nous souhaitons prédire, disons $(Y_t)_{1\leq t\leq T}$. À chaque instant $1 \leq t \leq T$ nous avons accès à toutes les observations $(X_t)_{1\leq t\leq T}$ et à un certain ensemble de prédicteurs $f_\Theta = \{f_\theta, \boldsymbol{\theta} \in \Theta\}$. Afin de faire leurs prédictions à l'instant t, les prédicteurs dans f_Θ peuvent avoir accès à l'échantillon précédent $(Y_s)_{s<t}$ mais pas nous. C'est une caractéristique intéressante de notre cadre : nous n'exploitons pas directement la suite $(Y_t)_{1\leq t\leq T}$ mais seulement à travers f_Θ. Pour un souci de simplicité nous admettons que les suites sont à valeurs réelles.
Soit ℓ la fonction perte. L'ensemble Θ est aussi indexé par, et possiblement changeant avec T. Soit $d_T > 0$ et supposons que pour tout $\boldsymbol{\theta} \in \Theta_T$ la fonction f_θ est définie sur $\mathbb{R}^{d_T}$, cela veut dire que pour tout $t > d_T$ la prédiction de Y_t qui correspond à $\boldsymbol{\theta}$ est donnée par $f_{\boldsymbol{\theta}}(\boldsymbol{Y}_{t-1:t-d_T})$. Nous équipons Θ_T avec la mesure prior π_T et construisons les prédicteurs de Gibbs

$$\hat{f}_{\eta,T}(\cdot|X) = \int_\Theta \nu_{\boldsymbol{\theta}}(\eta, T, X)\, f_{\boldsymbol{\theta}}(\cdot)\, \pi_T(\mathrm{d}\boldsymbol{\theta}) \ , \tag{1.6.1}$$

où les poids d'agrégation dépendent seulement du taux d'apprentissage η, de l'ensemble d'apprentissage X et de sa taille T. Le coefficient $\nu_{\boldsymbol{\theta}}$ satisfait le suivant (voir Alquier and Wintenberger (2012))

$$\nu_{\boldsymbol{\theta}}(\eta, T, X) \propto \exp\left(-\frac{\eta}{T - d_T} \sum_{t=d_T+1}^{T} \ell\left(f_{\boldsymbol{\theta}}\left(X_{t-1:t-d_T}\right), X_t\right)\right) , \tag{1.6.2}$$

et

$$\int_\Theta \nu_{\boldsymbol{\theta}}(\eta, T, X)\, \pi_T(\mathrm{d}\boldsymbol{\theta}) = 1 \ . \tag{1.6.3}$$

Sous des hypothèses souples sur les innovations qui génèrent le processus X, la collection de prédicteurs, la fonction perte ℓ, l'ensemble Θ_T et la mesure π_T, l'inégalité oracle PAC-Bayésienne suivante se tient pour tout $\varepsilon \in (0,1)$ avec P-probabilité au moins $1-\varepsilon$

$$R\left(\hat{f}_{\eta_T,T}(\cdot|X)\right) \leq \inf_{\theta\in\Theta_T} R(f_\theta) + \mathcal{E}\frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}}\log\left(\frac{1}{\varepsilon}\right), \tag{1.6.4}$$

où $\eta_T = \log T$ et la constante $\mathcal{E}$ sont explicitement calculées des hypothèses.
Cette inégalité s'applique au prédicteur agrégé exact $\hat{f}_{\eta_T,T}(\cdot|X)$. Dans la pratique, l'intégrale (1.6.1) est approximée numériquement par $\bar{f}_{\eta,T,n}(\cdot|X) = \sum_{i=0}^{n-1} f_{\theta_i}/n$ où $(\theta_i)_{i\geq 0}$ sont des instances une chaine de Markov $\Phi_{\eta,T}(X)$ qui a $\nu_\cdot(\eta,T,X)\pi_T$ comme unique mesure invariante. Cette chaine de Markov est typiquement construite en utilisant une méthode MCMC. L'algorithme Metropolis-Hastings en constitue un exemple.
Une chaine de Markov ajoute une seconde source d'aléa au processus de prédiction. Nous définissons $\nu_{\eta,T}$, une distribution de probabilité sur $(X, \Phi_{\eta,T}(X))$. Supposant que $\Phi_{\eta,T}(X)$ est géométriquement ergodique et sous les suppositions qui mènent à l'inégalité (1.6.4), nous montrons que pour tout $\varepsilon \in (0,1)$ et $n \geq M(T,\varepsilon)$, avec $\nu_{\eta_T,T}$-probabilité au moins $1-\varepsilon$ nous avons

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot|X)\right) \leq \inf_{\theta\in\Theta_T} R(f_\theta) + \left(\mathcal{E} + \frac{2}{\log 2} + 2\right)\frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}}\log\left(\frac{1}{\varepsilon}\right), \tag{1.6.5}$$

où $\eta_T = \log T$, $\mathcal{E}$ est le même de l'inégalité (1.6.4) et $M(T,\varepsilon)$ dépend en particulier, du taux de convergence de $\Phi_{\eta_T,T}(X)$ vers sa distribution invariante.
Observez que les termes de droite des inégalités (1.6.4) et (1.6.5) sont du même ordre. À notre connaissance, des bornes comme (1.6.5) n'ont pas été établies avant pour des procédures d'agrégation dans un contexte PAC-Bayésien quand Θ est potentiellement non fini.
Afin d'illustrer notre résultat, nous considérons le cas simple d'un processus auto-régressif d'ordre d stable et réel (comme dans l'Exemple 6, page 10) avec des innovations normalement distribuées et de variance 1. Soit $\ell(x,y) = |x-y|$, $d_T = \lfloor \log T \rfloor$, $\Theta_T \subset \mathbb{R}^{d_T}$ et $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}'\boldsymbol{x}$ pour tout $\boldsymbol{x} \in \mathbb{R}^{d_T}$. La définition précise de Θ_T et du prior π_T se trouvent dans la Section 3.5.
Le prédicteur agrégé est linéaire aussi et peut être exprimé comme $\hat{f}_{\eta,T}(\boldsymbol{x}|X) = \widehat{\boldsymbol{\theta}}'_{\eta,T}(X)\boldsymbol{x}$, où

$$\widehat{\boldsymbol{\theta}}_{\eta,T}(X) = \int_\Theta \nu_\theta(\eta,T,X)\,\boldsymbol{\theta}\,\pi_T(\mathrm{d}\boldsymbol{\theta}),$$

avec ν définie comme dans (1.6.2)-(1.6.3). Nous utilisons l'algorithme de Metropolis-Hastings afin d'approximer $\widehat{\boldsymbol{\theta}}_{\eta,T}(X)$ (les détails sont donnés dans la Section 3.5) par $\bar{\boldsymbol{\theta}}_{\eta,T,n}$. Soit γ_0 la variance du processus X. Pour un nombre d'itérations n supérieur à $M^*(T,\varepsilon) = 9\gamma_0^3 T^2 \exp(\gamma_0 T/16)/(2\pi\varepsilon^2\log^3 T)$ nous garantissons que la

borne (1.6.5) est atteinte. Cette borne supérieure pour $M(T, \varepsilon)$, possiblement pessimiste, rend prohibitif le calcule computationnel. Le prédicteur $\bar{\theta}'_{\eta_T,T,n}\boldsymbol{x}$ ($T = 2^{12}$ et $n = 1000$) montre un comportement pauvre dans nos expériences numériques.

1.6.2 Processus sous-linéaires non stationnaires et processus auto-régressifs variables dans le temps

Le Chapitre 4 présente un résultat général pour des processus sous-linéaires. Nous faisons une étude plus approfondie du cas particulier des processus TVAR.

Bornes d'agrégation pour des processus sous-linéaires :

Considérons la suite réelle $X = (X_t)_{t\in\mathbb{Z}}$ sous-linéaire par rapport au bruit $(Z_t)_{t\in\mathbb{Z}}$. Nous rappelons que

$$|X_t| \leq \sum_{j\in\mathbb{Z}} A_t(j)\, Z_{t-j} \ ,$$

où $(A_t(j))_{t,j\in\mathbb{Z}}$ sont des coefficients non négatifs qui satisfont

$$A_* := \sup_{t\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} A_t(j) < \infty \ .$$

Supposez que les X_ts arrivent au four et à mesure. À chaque instant $1 \leq t \leq T$ nous avons accès à $(X_s)_{1\leq s\leq t-1}$ et à un certain ensemble de prédicteurs $\{X_t^{(i)}, i = 1, \ldots, N\}$ et nous souhaitons construire notre propre prédicteur en ligne pour X_t.

En nous servant des inégalités oracle purement déterministes, dérivées de (Stoltz, 2011, Theorem 1.7) et (Catoni, 2004, Proposition 2.2.1), une borne uniforme de la norme ℓ_1 des coefficient sous-linéaires variables dans le temps, une hypothèse de type Lipschitz sur les prédicteurs et des conditions de moment sur le bruit apparaissant dans la représentation linéaire de X, nous obtenons les inégalités oracle suivantes.

(i) Considérons un bruit Z avec moment d'ordre 4 fini et soit $\widehat{X} = (\widehat{X}_t)_{1\leq t\leq T}$ qui désigne le prédicteur agrégé obtenu en utilisant les poids (1.4.4) avec $\eta \propto ((\log N)/T)^{1/2}$. Nous avons

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \inf_{\nu\in\mathcal{S}_N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + C_1 \left(\frac{\log N}{T}\right)^{1/2} , \qquad (1.6.6)$$

avec la constante C_1 qui peut être calculée des hypothèses.

(ii) Admettons que le bruit Z a un moment d'ordre p fini pour $p > 2$ et soit $\widehat{X} = (\widehat{X}_t)_{1\leq t\leq T}$ le prédicteur agrégé obtenu en utilisant les poids (1.4.5) et $\eta \propto ((\log N)/T)^{2/p}$. Nous avons

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \min_{1\leq i\leq N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + C_2 \left(\frac{\log N}{T}\right)^{1-2/p} , \qquad (1.6.7)$$

où C_2 est explicitement calculée des hypothèses.

(iii) Supposons que le bruit Z a un moment exponentiel fini et soit $\widehat{X} = (\widehat{X}_t)_{1\le t\le T}$ le prédicteur agrégé obtenu des poids (1.4.5) avec $\eta \propto (\log(T/(\log N)))^{-2}$. Donc, lorsque $(\log N)/T \to 0$ nous avons

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + C_3\frac{\log N}{T}\left(\log\left(\frac{T}{\log N}\right)\right)^2, \tag{1.6.8}$$

et la constante C_3 dépend des hypothèses.

Yang (2004) a proposé des bornes pour le regret du meilleur prédicteur dans un contexte de suites de variables aléatoires possiblement dépendantes. Un des ingrédients fondamentaux de cet article-là est que les prédicteurs sont supposés restant à une distance bornée des moyennes conditionnelles. L'Inégalité (1.6.8) est comparable avec un de ces résultats, mais nous l'obtenons sous des plus souples hypothèses.
Même si le cadre i.i.d. et le notre sont dissimilaires, nous présentons une courte comparaison de résultats dans les deux contextes. Concernant les bornes de régret convexe, Exemple 11 (Tsybakov (2003)) présente le meilleur reste résiduel possible quand les prédictions sont bornées. Il est (grosso modo) $(\log N/T)^{1/2}$ si N est beaucoup plus grand que $T^{1/2}$ et N/T quand N est plus petit que $T^{1/2}$. Par conséquence notre borne (1.6.6) coïncide seulement dans le cas où N est beaucoup plus large que $T^{1/2}$. Ceci-dit, quand N est plus petit que $T^{1/2}$, une procédure d'agrégation plus complexe permet d'obtenir une borne pour le regret convexe avec un terme résiduel de l'ordre de $N(\log T)^3/T$ (voir (4.9.7) page 125) si le bruit a un moment exponentiel fini. Par ailleurs, en imposant des conditions de moment d'ordre p sur le bruit et en appliquant une borne uniforme sur les prédicteurs, Audibert (2009) montre que le taux d'agrégation optimal est $(\log N/T)^{1-2/(p+2)}$, légèrement plus petit que notre $(\log N/T)^{1-2/p}$ in (1.6.7).

Bornes d'agrégation pour des processus TVAR :

Dans le contexte des processus TVAR (voir Exemple 3, page 6), $\beta > 0$, $\delta \in (0,1)$, $R > 0$, $\rho \in (0,1]$ et $\sigma_+ > 0$ définissent l'hyper-paramètre $\lambda = (\beta, R, \delta, \rho, \sigma_+)$ indexant la classe $\mathcal{M}_\lambda$. Dans la Section 4.3.2 nous présentons une borne inférieure pour le risque de prédiction minimax (2.3.19). Pour T suffisamment grand nous obtenons que

$$\inf_{\widehat{X}\in\mathcal{P}_T}\sup_{X\in\mathcal{M}_\lambda}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] - \frac{1}{T}\sum_{t=1}^{T}\sigma^2\left(\frac{t}{T}\right)\right\} \ge m_\lambda T^{-2\beta/(2\beta+1)}. \tag{1.6.9}$$

Ce taux coïncide avec celui du prédicteur construit à partir de l'estimateur NLMS avec une taille du pas de gradient bien choisie si $\beta \in (0,2]$. D'où, $T^{-2\beta/(2\beta+1)}$ est la vitesse minimax optimale pour des processus $\mathcal{M}_\lambda$ (au moins si $\beta \in (0,2]$).
Soit $\beta_0 \in (0,\infty]$ et soit $\{\widehat{X}^{(\beta)}, \beta \in (0,\beta_0)\}$ une collection de prédicteurs β-minimax en vitesse (δ, R, ρ et σ_+ étant fixés). Si $\beta_0 < \infty$ nous faisons $N = \lceil \log T \rceil$ et sélectionnons $\beta_i = (i-1)\beta_0/N$ pour $i = 1,\dots,N$. Autrement nous faisons $N = \lceil (\log T)^2 \rceil$ et $\beta_i = (i-1)\beta_0/N^{1/2}$ pour $i = 1,\dots,N$. Afin de construire $\widehat{X}$, nous agrégeons les prédicteurs $\{\widehat{X}^{(\beta_i)}, i = 1,\dots,N\}$, chacun desquels profite d'une vitesse de convergence minimax

ou optimal pour leurs respectives sur indexes β, en utilisant les poids (1.4.5) et le taux d'apprentissage choisi comme il suit.

(i) si le bruit Z présente un moment d'ordre p fini pour un $p > 2$ donné et $\beta_0 \leq (p-2)/4$, soit $\eta \propto (\log(\lceil \log T \rceil)/T)^{2/p}$,

(ii) si le bruit Z présente un moment exponentiel fini, soit $\eta \propto (\log T)^{-3}$.

Soit $\Lambda = \{(\beta, R, \delta, \rho, \sigma_+) : \beta \in (0, \beta_0)\}$. Nous montrons, avec l'aide des inégalités oracle énoncées précédemment, que $\widehat{X} = (\widehat{X}_{t,T})_{1 \leq t \leq T}$ est Λ minimax adaptatif, ceci signifie que

$$\sup_{X \in \mathcal{M}_\lambda} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] - \frac{1}{T} \sum_{t=1}^{T} \sigma^2 \left(\frac{t}{T} \right) \right\} \leq M_\lambda T^{-2\beta/(2\beta+1)} ,$$

pour tout $\boldsymbol{\lambda} \in \Lambda$.
Une caractéristique importante de $\widehat{X}$ est qu'il peut être calculé récursivement et qu'il est donc applicable dans un contexte de prédiction en ligne. L'algorithme suivant détaille la procédure en utilisant le NLMS, en supposant que $\beta_0 = 1$ et que le bruit a un moment exponentiel fini.

Algorithme 2: Prédiction adaptative en ligne

paramètres la valeur de T, l'ordre d;
initialisation $X_{s,T} = 0$ pour $-d \leq s \leq 0$, $\eta = (\log T)^{-3}$, $N = \lceil \log T \rceil$, $\widehat{\boldsymbol{\theta}}_{i,-1,T} = \mathbf{0} \in \mathbb{R}^d$ pour $i = 1, \ldots, N$, $t = 1$, $\widehat{\alpha}_t = (1/N)_{i=1,\ldots,N}$;
tant que l'entrée $X_{t-1,T}$ *est donnée;*
faire
  pour $i = 1$ **jusqu'à** N **faire**
    $\beta_i = (i-1)/N$;
    $\mu_i = T^{-2\beta_i/(2\beta_i+1)}$;
    $\widehat{\boldsymbol{\theta}}_{i,t-1,T} = \widehat{\boldsymbol{\theta}}_{i,t-2,T} + \mu_i \left(X_{t-1,T} - \widehat{\boldsymbol{\theta}}'_{i,t-2,T} \boldsymbol{X}_{t-2:t-d-1,T}\right) \dfrac{\boldsymbol{X}_{t-2:t-d-1,T}}{1 + \mu_i \left|\boldsymbol{X}_{t-2:t-d-1,T}\right|^2}$;
    pour $k = 1$ **jusqu'à** d **faire**
      $\widehat{\boldsymbol{\theta}}_{i,t-1,T}(k) = \min\left\{\max\left\{-\binom{n}{k}, \widehat{\boldsymbol{\theta}}_{i,t-1,T}(k)\right\}, \binom{n}{k}\right\}$;
    $\widehat{X}^{(i)}_{t,T} = \widehat{\boldsymbol{\theta}}'_{i,t-1,T} \boldsymbol{X}_{t-1:t-d,T}$;
  $\widehat{X}_{t,T} = \widehat{X}^{[\widehat{\alpha}_t]}_{t,T} = \sum_{i=1}^N \widehat{\alpha}_{i,t} \widehat{X}^{(i)}_{t,T}$;
  retourner $\widehat{X}_{t,T}$;
  $t = t + 1$;
  tant que l'entrée $X_{t-1,T}$ *est donnée;*
  faire
    pour $i = 1$ **jusqu'à** N **faire**
      $v_{i,t} = \widehat{\alpha}_{i,t-1} \exp\left(-\eta \left(\widehat{X}^{(i)}_{t-1,T} - X_{t-1,T}\right)^2\right)$;
    $\widehat{\alpha}_t = \left(v_{i,t} / \sum_{k=1}^N v_{k,t}\right)_{i=1,\ldots,N}$;

1.6.3 Processus localement stationnaires

Soit $d \in \mathbb{N}^*, \beta \geq 2, R, f_- > 0$ et $\boldsymbol{\lambda} = (\beta, R, f_-)$. Dans le Chapitre 5 notre étude regarde $\Lambda'_1(\beta, R)$, un sous-ensemble de $\Lambda_1(\beta, R)$ (voir Section 5.2.2). Considérons $\mathcal{M}_{\boldsymbol{\lambda}}$ l'ensemble de tous les processus localement stationnaires par rapport à la Définition 14 (ce généralise ceux caractérises par Équation (1.2.5)), dont les densités spectrales $f(\cdot, \omega) \in \Lambda'_1(\beta, R)$ pour tout ω et $f \geq f_-$ (voir Définition 13). Nous abordons le problème de régression suivant :

$$\boldsymbol{\theta}^*_{t,T} = \underset{\boldsymbol{\theta}=[\theta_1 \ldots \theta_d] \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[\left(X_{t,T} - \sum_{k=1}^d \theta_k\, X_{t-k,T}\right)^2\right] = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}\left[\left(X_{t,T} - \boldsymbol{\theta}' \boldsymbol{X}_{t-1:t-d,T}\right)^2\right],$$

où $X \in \mathcal{M}_{\boldsymbol{\lambda}}$ et B' dénote la transposée de la matrice B.
Le vecteur $\boldsymbol{\theta}^*_{t,T}$ coïncide avec $\boldsymbol{\theta}(t/T)$ dans le cas des processus TVAR localement stationnaires.

Étant donnée $h : [0,1] \to \mathbb{R}$ et $m \leq T$, la fonction de covariance empirique $\widehat{\gamma}_m$ est définie en $\mathbb{R} \times \mathbb{Z}$ comme

$$\widehat{\gamma}_m(u,\ell) = \frac{1}{H_m} \sum_{\substack{t_1,t_2=1 \\ t_1-t_2=\ell}}^{m} h\left(\frac{t_1}{m}\right) h\left(\frac{t_2}{m}\right) X_{\lfloor uT \rfloor + t_1 - m/2, T} X_{\lfloor uT \rfloor + t_2 - m/2, T} \,, \tag{1.6.10}$$

où $H_m = \sum_{k=1}^{m} h^2(k/m)$.
En nous appuyant sur la définition précédente nous proposons un estimateur $\widetilde{\boldsymbol{\theta}}_{t,T}$ pour $\boldsymbol{\theta}^*_{t,T}$ en deux pas. Soit $k = \lceil \beta \rceil - 1$ et $M \in 2^{k+1}\,\mathbb{N}^*$. Dans un premier pas nous utilisons les équations de Yule-Walker (voir Dahlhaus and Giraitis (1998)) et pour $m \in \{M/2^j, j = 0, \ldots, k\}$ nous construisons

$$\widehat{\boldsymbol{\theta}}_{t,T}(m) \;=\; \widehat{\Gamma}^{-1}_{t,T,m} \widehat{\boldsymbol{\gamma}}_{t,T,m} \,,$$

où $\widehat{\boldsymbol{\gamma}}_{t,T,m} = [\widehat{\gamma}_m(t/T, 1) \;\ldots\; \widehat{\gamma}_m(t/T, d)]'$, $\widehat{\Gamma}_{t,T,m}$ est la matrice de covariances empiriques $\widehat{\Gamma}_{t,T,m} = (\widehat{\gamma}_m(t/T, i-j); i, j = 1, \ldots, k)$ et $\widehat{\gamma}_m$ est la fonction de covariance empirique comme définie dans (1.6.10).
Le deuxième pas consiste à combiner tous les $\widehat{\boldsymbol{\theta}}_{t,T}(m)$ pour $m \in \{M/2^j, j = 0, \ldots, k\}$ de la manière suivante. Soit $\boldsymbol{\alpha} = [\alpha_0 \;\ldots\; \alpha_k]' \in \mathbb{R}^{k+1}$ la solution de l'équation $A\boldsymbol{\alpha} = \boldsymbol{e}_1$ où A est la matrice réelle de dimension $(k+1) \times (k+1)$ avec les entrées $A_{i,j} = 2^{-(i-1)(j-1)}$ et $\boldsymbol{e}_1 = [1\; 0\; \ldots\; 0]' \in \mathbb{R}^{k+1}$ contient un 1 dans sa première composante et zéro partout ailleurs. Définissons ensuite $\widetilde{\boldsymbol{\theta}}_{t,T} = \sum_{j=0}^{k} \alpha_j \widehat{\boldsymbol{\theta}}_{t,T}(M/2^j)$.
Dénotons $\widehat{X}_{t,T} = \widetilde{\boldsymbol{\theta}}'_{t,T} \boldsymbol{X}_{t-1:t-d,T}$, $\widehat{X}^*_{d,t,T} = (\boldsymbol{\theta}^*_{t,T})' \boldsymbol{X}_{t-1:t-d,T}$ et $\widehat{X}^*_{t,T} = \mathbb{E}[X_{t,T} | \sigma(X_{s,T}, s \leq t-1)]$. Nous obtenons que, pour T suffisamment grand et $q > 0$

$$\sup_{X \in \mathcal{M}_\lambda} \mathbb{E}\left[\left\| \widetilde{\boldsymbol{\theta}}_{t,T}(M) - \boldsymbol{\theta}^*_{t,T} \right\|^q \right] \leq C_1 \left(\frac{1}{M^{1/2}} + \left(\frac{M}{T} \right)^{\beta} \right)^q \,,$$

et

$$\mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] \leq \mathbb{E}\left[\left(\widehat{X}^*_{t,T} - X_{t,T} \right)^2 \right] + \mathbb{E}\left[\left(\widehat{X}^*_{t,T} - \widehat{X}^*_{d,t,T} \right)^2 \right] + C_2 \left(\frac{1}{M^{1/2}} + \left(\frac{M}{T} \right)^{\beta} \right)^2 + C_3 \left(\frac{1}{M^{1/2}} + \left(\frac{M}{T} \right)^{\beta} \right) \left(\mathbb{E}\left[\left(\widehat{X}^*_{t,T} - \widehat{X}^*_{d,t,T} \right)^2 \right] \right)^{1/2} \,,$$

où C_1, C_2 et C_3 dépendent seulement de $\boldsymbol{\lambda}$.
Le résultat peut être appliqué aux processus TVAR localement stationnaires générés par $\boldsymbol{\theta} \in s_d(\delta) \cap \Lambda'_d(\beta, R)$ (avec $\delta \in (0,1)$) et $\sigma \in \Lambda'(\beta, R) \cap [\rho\sigma_+, \sigma_+]^{(-\infty,1]}$. Dans ce premier cas nous définissons $\boldsymbol{\lambda} = (\beta, R, \delta, \rho, \sigma_+)$ comme dans (2.3.17). En spécifiant $M \propto T^{-2\beta/(2\beta+1)}$ nous obtenons la vitesse minimax pour le regret

$$\sup_{X \in \mathcal{M}_\lambda} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] - \frac{1}{T} \sum_{t=1}^{T} \sigma^2\left(\frac{t}{T} \right) \right\} \leq M_\lambda T^{-2\beta/(2\beta+1)} \,.$$

La procédure fondée sur l'algorithme NLMS qui cherche à estimer $\boldsymbol{\theta}$ étudiée par Moulines et al. (2005) garantit des vitesses de convergence minimax pour $\beta \in (0, 2]$ dans le modèle TVAR. Nous ne sommes au courant d'aucun résultat similaire pour $\beta > 2$.

1.7 PERSPECTIVES

Agréger un nombre infini de prédicteurs peut poser le problème de la calculabilité comme évoqué dans le Chapitre 3. Lors que nous utilisons des méthodes de Monte Carlo par chaines de Markov, $M(T, \varepsilon)$, le nombre d'itérations nécessaires pour atteindre une précision numérique du même ordre que le risque de prédiction peut exploser avec la valeur de T (l'horizon). Dans ce contexte, des algorithmes capables de passer à l'échelle à coût computationnel raisonnable devront être examinés.

En utilisant un nombre fini de prédicteurs, dans le Chapitre 4 nous étudions des bornes supérieures pour le regret convexe et pour celui lié au meilleur prédicteur pour les processus sous-linéaires. Dans le cas précis où le bruit associé au processus tout comme les prédicteurs ont des moments d'ordre p finis, l'optimalité de ces bornes reste un problème ouvert.

Une analyse détaillée des stratégies qui ne reposent pas sur une information a priori sur le processus et ses prédicteurs doit être menée. Il semblerait possible que nous puisons obtenir une telle amélioration sans faire ralentir nos bornes oracle.

Étant donné que les processus TVAR sont en même temps sous-linéaires et localement stationnaires, l'analyse des chapitres 4 et 5 leur est applicable. Si l'ordre d est connu, notre contribution permet de proposer un prédicteur minimax adaptatif en utilisant les algorithmes NLMS, Yule-Walker et l'agrégation. Par contre, quand d n'est pas connu, il n'est pas clair commet sélectionner les prédicteurs à agréger.

Lors que nous travaillons avec des processus non stationnaires, il est intéressant de séparer les rôles du nombre d'observations ou horizon (que nous appelons T) et celui de la fréquence échantillonnage ω (supposée être T^{-1} tout au long de cette thèse). Une traduction de nos hypothèses et résultats pour des séries temporelles exprimées comme $(X_{t,T})_{1\leq t\leq T}$ à des séries temporelles exprimées comme $(X_{t,\omega})_{t\geq 1}$ pourrait être un premier pas dans cette direction.

2 Introduction

2.1 Content and notation

The present chapter lays the groundwork for our research. In Section 2.2, we present the models we are interested in: some classes of weakly dependent processes and of locally stationary processes. Our ultimate objective is to propose efficient forecasting methods on these models. The quality of a prediction is measured by a loss function. We need it to be as small as possible, typically in expectation or with high probability. These notions are formalized in a general framework introduced in Section 2.3.1. In Section 2.3.2 we explain the optimality features associated with the prediction algorithms we explore. The exponentially weighted aggregation is the cornerstone of this thesis, we give a brief overview about it in Section 2.4. In Section 2.5 we enumerate the precise problems that we tackle and in Section 2.6 we present our main results. Section 2.7 advances possible research directions in the continuity of our work.
Throughout this chapter, for $\boldsymbol{a} \in \mathbb{R}^q$ with $q \in \mathbb{N}^*$, $\|\boldsymbol{a}\|$ denotes its Euclidean norm, $\|\boldsymbol{a}\| = (\sum_{i=1}^q a_i^2)^{1/2}$ and $\|\boldsymbol{a}\|_1$ its ℓ_1 norm $\|\boldsymbol{a}\|_1 = \sum_{i=1}^q |a_i|$. Bold characters represent column vectors and normal characters their components as in $\boldsymbol{y} = (y_i)_{i\in\mathbb{Z}}$. The use of subscripts with colon ':' refers to a subvector of consecutive components $\boldsymbol{y}_{1:k} = [y_1 \dots y_k]'$ (forward), $\boldsymbol{y}_{k:1} = [y_k \dots y_1]'$ (backward) or elements of a sequence $\boldsymbol{X}_{1:k} = [X_1 \dots X_k]'$ (forward), $\boldsymbol{X}_{k:1} = [X_k \dots X_1]'$ (backward); in all cases they are k dimensional vectors.

2.2 Models

Independent and identically distributed random variables are the prima materia of a wide part of the statistical literature. Although the present contribution relies on them, they are not our main target. The particular problems that we study are focused on sequences of random variables which may be (and it is interesting when they are) dependent and may have (and it is interesting when they have) a distribution that evolves. The next two subsections briefly put the models that we study into context.

2.2.1 Weak dependence

The weak dependence paradigm, proposed by Doukhan and Louhichi (1999), is an approach that makes explicit the asymptotic independence between two distanced

moments in a time series. It represents a unifying viewpoint of other competitive notions such as mixing conditions, more adapted to σ-fields. The α (strong) and the β-mixing coefficients for example, were introduced by Rosenblatt (1956) and Volkonskiĭ and Rozanov (1959) respectively. We revisit a couple of definitions that are crucial in our investigation. We refer to Dedecker et al. (2007) for a comprehensive material in weak dependence.
Recall that two random variables X and Y defined on the same probability space are said to be independent if and only if $\mathrm{cov}(f(X), g(Y)) = 0$ for all real Borel-measurable and bounded functions f and g. A relaxation of the independence is the following.

Definition 1 (Weak dependence)**.** *The sequence $(X_t)_{t\in\mathbb{Z}}$ with values in a locally compact topological space $\mathcal{X}$ (typically $\mathbb{R}^d$) is said to be weakly dependent, if there exists a class $\mathcal{F}$ of functions such that for any $u, v \in \mathbb{N}^*$ and any $f, g \in \mathcal{F}$ respectively defined in $\mathcal{X}^u$ and $\mathcal{X}^v$ the following asymptotic relation holds*

$$\varepsilon(r) = \sup_{i_1\le\ldots\le i_u<i_u+r\le j_1\le\ldots\le j_v} \left|\mathrm{cov}\left(f(X_{i_1},\ldots,X_{i_u}), g\left(X_{j_1},\ldots,X_{j_v}\right)\right)\right| \to 0\,, \text{ when } r\to\infty\,.$$

Bernoulli shifts is a very rich class of weakly dependent processes. It is the first model that we study.

Definition 2 (Bernoulli shift)**.** *Let $(\xi_t)_{t\in\mathbb{Z}}$ be a sequence of independent real-valued random variables and $H : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ be a Borel function. A Bernoulli shift is a sequence $(X_t)_{t\in\mathbb{Z}}$ satisfying*

$$X_t = H\left(\xi_{t-j}, j\in\mathbb{Z}\right)\,. \tag{2.2.1}$$

Not all sequences $(\xi_t)_{t\in\mathbb{Z}}$ and measurable functions $H : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ define a Bernoulli shift. Convergence issues may arise from the expression of H and the particularities of ξ_t. For illustrating this ambiguity let ξ_t be a uniform random variable in $[1, 2]$ for all t and $H(\boldsymbol{u}) = \sum_{j\in\mathbb{Z}}(-1)^j u_j$. In this case the representation (2.2.1) makes no sense.
The expression (2.2.1) is well defined when $(\xi_t)_{t\in\mathbb{Z}}$ have uniformly bounded absolute moments and H is Lipschitz, that is if $\sup_{t\in\mathbb{Z}}\mathbb{E}[|\xi_t|] < \infty$ and for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathbb{Z}}$

$$|H(\boldsymbol{u}) - H(\boldsymbol{v})| \le \sum_{j\in\mathbb{Z}} A_j \left|u_j - v_j\right|\,, \tag{2.2.2}$$

with

$$A_* = \sum_{j\in\mathbb{Z}} A_j < \infty\,. \tag{2.2.3}$$

Considering $\mathcal{F} = \cup_{j\ge1}\mathcal{F}_j$ in Definition 1, where $\mathcal{F}_j$ is the set of bounded Lipschitz functions from $\mathbb{R}^j$ to $\mathbb{R}$, we can show that the previously well defined Bernoulli shift

(satisfying (2.2.1), (2.2.2) and (2.2.3)) is weakly dependent (see (Doukhan and Louhichi, 1999, Lemma 9) and (Dedecker et al., 2007, Lemma 3.1)).
If for all $t \in \mathbb{Z}$ the instance X_t only depends on $(\xi_s)_{s \leq t}$, that is

$$X_t = H\left(\xi_{t-j}, j \geq 0\right) ,$$

we say that the process $(X_t)_{t\in\mathbb{Z}}$ is a Causal Bernoulli Shift (CBS) and the real random variables $(\xi_t)_{t\in\mathbb{Z}}$ are called innovations.
Bernoulli shifts regroup several weakly dependent processes derived from stationary sequences. They also provide examples of processes that are weakly dependent, but not mixing (see Rosenblatt (1980)). In the following we present two examples of Bernoulli shifts.

Example 1 (Infinite Moving Average (MA(∞)) process). Let $(\xi_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. real random variables, centred and with variance 1. The MA(∞) process is defined by the representation

$$X_t = \sum_{j\in\mathbb{Z}} a_j \xi_{t-j} ,$$

where $\sum_{j\in\mathbb{Z}} |a_j| < \infty$.

Volterra processes are a generalisation of MA(∞) processes of Example 1 (see (Doukhan, 2003, Section 2.4)).

Example 2 (Volterra process). Let $(\xi_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. real random variables, centred and with variance 1 and let $v_0 \in \mathbb{R}$. Consider the sequence $(a_{k;i_1,\ldots,i_k})_{k\in\mathbb{N}^*,(i_1,\ldots,i_k)\in\mathbb{Z}^k}$ of real numbers. Set

$$\begin{aligned} X_t &= v_0 + \sum_{k=1}^{\infty} V_{k,t} , \\ V_{k,t} &= \sum_{i_1<\ldots<i_k} a_{k;i_1,\ldots,i_k} \prod_{j=1}^{k} \xi_{t-i_j} . \end{aligned}$$

Observe that if the coefficients satisfy

$$\sum_{k=1}^{\infty} \sum_{i_1<\ldots<i_k} \left|a_{k;i_1,\ldots,i_k}\right| < \infty ,$$

then, the Minkowski inequality implies that $X_t \in L^2$ for all $t \in \mathbb{Z}$.

We end the present section by giving a useful result.

A concentration inequality :
On weakly stationary process, as in the more classical i.i.d. setting, the proof of results concerning the quality of the prediction involve the use of concentration inequalities (see Massart (2007)). For the sake of completeness, we provide a Hoeffding type exponential inequality (see (Rio, 2000, Theorem 1) and (Alquier and Wintenberger, 2012, Proposition 4.2)) satisfied by the CBS.
Let $n > 0$. We say that $f : \mathbb{R}^n \to \mathbb{R}$ is an M-Lipschitz function if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq M \|\boldsymbol{x} - \boldsymbol{y}\|_1 \ .$$

2

Theorem 2.2.1. *Let $(X_t)_{t\in\mathbb{Z}}$ be a bounded CBS associated with the bounded innovations $(\xi_t)_{t\in\mathbb{Z}}$, let $f : \mathbb{R}^n \to \mathbb{R}$ be a 1-Lipschitz function and $x > 0$. The following inequality holds*

$$\mathbb{P}\left(f(\boldsymbol{X}_{1:n}) - \mathbb{E}\left[f(\boldsymbol{X}_{1:n})\right] \geq x\right) \leq \exp\left(-x^2\left[2n\left(\|X_0\|_\infty + 2\sum_{j\geq 0} jA_j \|\xi_0\|_\infty\right)\right]^{-1}\right),$$

where the sequence $(A_j)_{j\geq 0}$ fulfills Inequality (2.2.2), and $\|X_0\|_\infty$ and $\|\xi_0\|_\infty$ are the respective supreme values of $|X_0|$ and $|\xi_0|$.

2.2.2 Local stationarity

A particular property of stationary process is that they involve a collection of identically distributed random variables. A less restrictive class is that of weakly stationary processes (also called second-order stationary processes). We recall the definition (see (Brockwell and Davis, 2002, Section 1.4 and Theorem 2.1.1)).

Definition 3 (Weakly stationary process)**.** *Let $\mu \in \mathbb{R}$ and $\gamma : \mathbb{Z} \to \mathbb{R}$ be a symmetric and non-negative definite function. We say that a real-valued process $(X_t)_{t\in Z}$ is weakly stationary if the three following conditions are satisfied.*

(i) $\mathbb{E}|X_t|^2 < \infty$ *for all* $t \in \mathbb{Z}$,

(ii) $\mathbb{E}[X_t] = \mu$ *for all* $t \in \mathbb{Z}$,

(iii) $\mathrm{Cov}(X_s, X_t) = \gamma(s - t)$ *for all* $s, t \in \mathbb{Z}$.

A crucial characteristic of weakly stationary processes is that their spectra is constant. On the other hand, several studies about time series having an evolving spectrum (among other characteristics) appeared in the second half of the last century (see for example Granger (1964) and Priestley (1965)). Three decades later, Dahlhaus (1996b) introduced an approach allowing for fruitful local asymptotic considerations.
Suppose for example that the observations correspond to the model

$$X_t = \theta_t X_{t-1} + \sigma_t \xi_t \ . \tag{2.2.4}$$

At first sight, we may wonder how an estimator of the function θ, obtained from the sample $(X_t)_{1\leq t\leq T}$, behaves when T is large enough. To consider this kind of questions is usually contradictory to the non-stationarity assumption. As the probabilistic structure of the process may substantially differ from small to larger values of t, the information coming with new observations (let us say t large enough) may be useless to estimate θ_t for small values of t.
To cope with this difficulty, Dahlhaus (1996b) came up with the idea of locally stationary processes, which admit the representation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} \exp\left(it\omega\right) A^0_{t,T}\left(\omega\right) \xi\left(\mathrm{d}\omega\right) \ , \tag{2.2.5}$$

where, in particular, there exist a constant K and a (unique) $2\pi-$ periodic function $A : (-\infty, 1] \times \mathbb{R} \to \mathbb{C}$ with $A(u, -\omega) = \overline{A(u, \omega)}$ such that for all T

$$\sup_{t,\omega} \left|A^0_{t,T}\left(\omega\right) - A\left(\frac{t}{T}, \omega\right)\right| \leq \frac{K}{T} \ . \tag{2.2.6}$$

The artificial introduction of the dependence on the horizon T and the additional assumptions on $(X_t)_{1\leq t\leq T}$, open the door to meaningful asymptotic statistical procedures. This locally stationary model covers essentially time varying linear processes. A considerable part of our work deals with it.
In the case of the process described by Equation (2.2.4), for instance, it is locally stationary when the sequence $(\theta_{t,T})_{1\leq t\leq T}$ fulfills

$$\sup_{T\geq 1} \sum_{t=1}^{T} \left|\theta_{t,T} - \theta\left(\frac{t}{T}\right)\right| < \infty \ , \tag{2.2.7}$$

where $\theta : [0, 1] \to \mathbb{R}$ is a suitable function (see Dahlhaus (2009) and the references therein).

Example 3 (TVAR model). A special version of the model (2.2.4) is the time varying autoregressive (TVAR) process. It satisfies the recursive equation

$$X_{t,T} = \sum_{j=1}^{d} \theta_j\left(\frac{t}{T}\right) X_{t-j,T} + \sigma\left(\frac{t}{T}\right)\xi_t \ ,$$

where $(\xi_t)_{t\in\mathbb{Z}}$ is a white noise process and $\boldsymbol{\theta} = [\theta_1 \ \ldots \ \theta_d] \in s_d(\delta)$ with $\delta \in (0, 1)$.
The set $s_d(\delta)$ is associated with the stability of $(X_{t,T})_{1\leq t\leq T}$ and is defined as

$$s_d\left(\delta\right) = \left\{\boldsymbol{\theta} : (-\infty, 1] \to \mathbb{R}^d, 1 - \sum_{j=1}^{d} \theta_j\left(u\right) z^j \neq 0, \forall\, |z| < \delta^{-1}, u \in [0, 1]\right\} \ , \tag{2.2.8}$$

(see (Dahlhaus, 1996b, Theorem 2.3)).
Regularity conditions on $\boldsymbol{\theta}$ are necessary to obtain interesting statistical results for TVAR processes. Requiring derivatives up to certain order is a quite standard one (see (Dahlhaus and Giraitis, 1998, Assumption 2.1 (i) and Assumption 3.1 (ii)-(iii))). Moulines et al. (2005) rely in a more flexible assumption compared to Dahlhaus and Giraitis (1998): let $R, \beta > 0$ and let k be the biggest integer strictly smaller than β; in addition to $\boldsymbol{\theta} \in s_d(\delta)$, they suppose that $\boldsymbol{\theta} \in \Lambda_d(\beta, R)$, where

$$\Lambda_d(\beta, R) = \left\{ \boldsymbol{\theta} \in C^k\left((-\infty, 1], \mathbb{R}^d\right) : \sup_{0<|s-s'|<1} \frac{\left|\boldsymbol{\theta}^{(k)}(s) - \boldsymbol{\theta}^{(k)}(s')\right|}{|s-s'|^{\beta-k}} \leq R \right\} . \tag{2.2.9}$$

Similar ideas of local stationarity were developed in different contexts. Time varying autoregressive conditional heteroscedastic (tvARCH) processes (see (Dahlhaus and Subba Rao, 2006, Section 2)) and generalised autoregressive conditional heteroscedastic processes (tvGARCH) (see (Subba Rao, 2006, Section 5)) are examples of it. A common ingredient to these approaches is that the process can be locally approximated by its stationary version.
In the following we introduce a simple extension of the locally stationary linear model. Let $(Z_t)_{t\in\mathbb{Z}}$ be a sequence of non-negative random variables (not necessarily i.i.d.). The process $(X_t)_{t\in\mathbb{Z}}$ is said to be sub-linear with respect to $(Z_t)_{t\in\mathbb{Z}}$ if

$$|X_t| \leq \sum_{j\in\mathbb{Z}} A_t(j)\, Z_{t-j} , \tag{2.2.10}$$

where $(A_t(j))_{t,j\in\mathbb{Z}}$ are non-negative coefficients satisfying

$$A_* := \sup_{t\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} A_t(j) < \infty .$$

Results concerning $(X_t)_{t\in\mathbb{Z}}$ are deduced by imposing additional assumptions on $(Z_t)_{t\in\mathbb{Z}}$. For example, the Minkowski inequality implies the existence of moments of order p for the process $(X_t)_{t\in\mathbb{Z}}$ when the moments of order p of $(Z_t)_{t\in\mathbb{Z}}$ are uniformly bounded.

Example 4 (A non-linear model). A non-linear example of sub-linear processes is

$$X_t = g_t(X_{t-1}) + \xi_t ,$$

where the $(\xi_t)_{t\in\mathbb{Z}}$ are i.i.d. and $(g_t)_{t\in\mathbb{Z}}$ is a time varying sequence of sub-linear functions fulfilling, for all t

$$|g_t(x)| \leq \alpha|x| ,$$

for some $\alpha \in (0, 1)$. Then, we have that

$$|X_t| \leq \alpha|X_{t-1}| + |\xi_t| .$$

Iterating this equation backwards yields (2.2.10) with $Z_t = |\xi_t|$ and $A_t(j) = \alpha^j$. In the stationary framework, where $g = g_t$ does not depend on t, a well understood illustration of a non-linear case is given by the threshold autoregressive model where g is piecewise linear, see Tong and Lim (1980).

2.3 PREDICTION

2.3.1 General setting

In this section we present a general framework that includes a considerable number of prediction problems investigated in the literature. Consider a time series $(Z_t)_{1\le t\le T}$. The construction of one step ahead predictors $(\widehat{Z}_t)_{1\le t\le T}$ sometimes relies on a learning data set. A typical case is when we split the data into training set and validation set. The predictors learn exclusively from the training set, while the validation set is used to evaluate the quality of the prediction. In the following, we provide the notions required to construct our formalism.
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{H} \subset \mathcal{F}$ be a sub σ-field that we call the learning σ-field, and let $(\mathcal{F}_t)_{t\ge 0}$ be a filtration that we name the predicting filtration. The σ-field $\mathcal{H}$ contains the information from the learning data set. It can be reduced to the trivial σ-field if none is available.
Consider a $\mathcal{Z}$-valued process $(Z_t)_{t\ge 0}$, adapted to $(\mathcal{F}_t)_{t\ge 0}$, where $(\mathcal{Z}, \ell)$ is a metric space.

Definition 4 (Predictor)**.** *For all $t \ge 1$, we say that $\widehat{Z}_t$ is a predictor of Z_t if it is measurable with respect to the joint σ-field $\mathcal{H} \vee \mathcal{F}_{t-1}$.*
For any $T \ge 1$, we denote by $\mathcal{P}_T$ the set of sequences $\widehat{Z} = (\widehat{Z}_t)_{1\le t\le T}$ of predictors of $(Z_t)_{1\le t\le T}$, that is, the set of all processes $\widehat{Z} = (\widehat{Z}_t)_{1\le t\le T}$ adapted to $(\mathcal{H} \vee \mathcal{F}_{t-1})_{1\le t\le T}$.

We define the learning loss as

$$\frac{1}{T}\sum_{t=1}^{T} \ell\left(\widehat{Z}_t, Z_t\right) . \tag{2.3.1}$$

The prediction risk is provided by the conditional expectation of the learning loss given the learning σ-field.

$$R_T\left(\widehat{Z}\middle|\mathcal{H}\right) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell\left(\widehat{Z}_t, Z_t\right)\middle|\mathcal{H}\right] . \tag{2.3.2}$$

The risk is defined as the expectation of the learning loss.

$$R_T\left(\widehat{Z}\right) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell\left(\widehat{Z}_t, Z_t\right)\right] . \tag{2.3.3}$$

Bear in mind that, depending on the context, a sequence of predictors $\widehat{Z}$ is more efficient when (2.3.1), (2.3.2) or (2.3.3) are smaller. Often, in the data splitting framework one seeks to minimize (2.3.2) with high probability. On the other hand, if $\mathcal{H}$ is the trivial σ-field (2.3.2) and (2.3.3) coincide and one looks for minimizing them.
The next two sections explain how classical prediction problems connected to this thesis fit into our framework. They also provide standard results.

2.3.1.1 Predicting a dependent process given a learning data set

Suppose that we observe the first T instances of a possibly dependent stochastic process $X = (X_t)_{t\geq 1}$ lying on $\mathcal{X}$. The distribution of the whole process is denoted by P. Suppose moreover that we want to predict the first T instances of another stochastic process $Y = (Y_t)_{t\geq 1}$, independent of X, with values in $\mathcal{X}$ and distributed also according to P.
This set-up commonly arises when we split the data into training set and validation set. We refer to Audibert and Catoni (2010, 2011); Hsu et al. (2011) when all the observations are independent. A more complex situation is when the available data are dependent (like in the case of an AR(d)). Even though the premise of independence is not fulfilled, in practice, we may split the observations and hope that the predictors built from the training set (X) lead to a low risk on the validation set (Y). For theoretical purposes, it is convenient to assume that X and Y are independent, although it may not be true in practice.
Let $\mathcal{H} = \sigma(X_t, 1 \leq t \leq T)$ be the learning σ-field and let $\mathcal{F} = (\mathcal{F}_t)_{t\geq 1}$ be the natural filtration associated with Y, where $\mathcal{F}_t = \sigma(Y_s, 1 \leq s \leq t)$. In this context, we construct for each $t = 1, \dots, T$ an application $\hat{f}_t : \mathcal{X}^T \to \mathcal{X}^{\mathcal{X}^{t-1}}$. We denote by $\hat{f}_t(\cdot|\boldsymbol{X}_{1:T})$ the function that predicts Y_t from $Y_1, \dots, Y_{t-1}$, given the observations $X_1, \dots, X_T$. Then, set $\widehat{Y}_t = \hat{f}_t(\boldsymbol{Y}_{1:t-1}|\boldsymbol{X}_{1:T})$.
The following expression corresponds to the prediction risk defined by Equation (2.3.2)

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell\left(\hat{f}_t\left(\boldsymbol{Y}_{1:t-1}|\,\boldsymbol{X}_{1:T}\right), Y_t\right)\middle|\,\boldsymbol{X}_{1:T}\right] = \frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{X}^{\mathbb{N}^*}}\ell\left(\hat{f}_t\left(\boldsymbol{y}_{1:t-1}|\boldsymbol{X}_{1:T}\right), y_t\right)P\left(\mathrm{d}\boldsymbol{y}\right) \ . \quad (2.3.4)$$

Let us consider the case where the process is stationary and the predictor $\hat{f}_t(\cdot|\boldsymbol{X}_{1:T})$ only exploits $Y_{t-d}, \dots, Y_{t-1}$ for a fixed $d > 0$. Moreover, we assume that $\hat{f}_t$ does not depend on t. Dropping off the first d terms of (2.3.4) leads to analyse

$$R\left(\widehat{Y}\middle|\,\mathcal{H}\right) = \int_{\mathcal{X}^{\mathbb{N}^*}}\ell\left(\hat{f}\left(\boldsymbol{y}_{1:d}|\boldsymbol{X}_{1:T}\right), y_{d+1}\right)P\left(\mathrm{d}\boldsymbol{y}\right) \ ,$$

which no longer depends on T.
The corresponding risk takes the form

$$\mathbb{E}\left[\ell\left(\hat{f}\left(\boldsymbol{Y}_{1:d}|\boldsymbol{X}_{1:T}\right), Y_{d+1}\right)\right] = \int_{\mathcal{X}^{\mathbb{N}^*}}\int_{\mathcal{X}^{\mathbb{N}^*}}\ell\left(\hat{f}\left(\boldsymbol{y}_{1:d}|\boldsymbol{x}_{1:T}\right), y_{d+1}\right)P\left(\mathrm{d}\boldsymbol{y}\right)P\left(\mathrm{d}\boldsymbol{x}\right) \ .$$

Example 5 (AR(∞)). Consider the observations $X_1, \dots, X_T$, of a real autoregressive process $X = (X_t)_{t\in\mathbb{Z}}$. It is defined by the recursive equation

$$X_t = \sum_{j=1}^{\infty} \theta_j X_{t-j} + \sigma\xi_t \ , \tag{2.3.5}$$

where $\boldsymbol{\theta} = [\theta_1 \ \dots]' \in \mathbb{R}^{\mathbb{N}^*}$, $|\boldsymbol{\theta}|_1 = \sum_{j=1}^{\infty} |\theta_j| < \infty$, $1 - \sum_{j=1}^{\infty} \theta_j z^j \neq 0$ for $|z| \leq 1$, $\sigma > 0$, and $(\xi_t)_{t\in\mathbb{Z}}$ is a sequence of i.i.d. centred random variables with variance 1.

Linear prediction of AR(∞) :

The standard approach for forecasting this class of processes dates back to Akaike (1969) (see extensions and generalisations in Berk (1974), Bhansali (1978) and Lewis and Reinsel (1985)). It consists in, for a fixed $d \in \mathbb{N}^*$, regressing X_t onto $X_{t-1}, \dots, X_{t-d}$. In other words, we build a predictor $\widehat{X}_t = \widehat{\boldsymbol{\theta}}' \boldsymbol{X}_{t-1:t-d}$, where $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1 \ \dots \ \widehat{\theta}_d]' \in \mathbb{R}^d$. Provided $d \in \mathbb{N}^*$, the estimator $\widehat{\boldsymbol{\theta}}$ is supposed to realize the infimum of the mean square prediction error

$$\frac{1}{T} \sum_{t=1}^{T} \left(X_t - \boldsymbol{\theta}' \boldsymbol{X}_{t-1:t-d} \right)^2 \ .$$

Hence, it is given by the Yule-Walker equations

$$\widehat{\boldsymbol{\theta}} \ = \ \widehat{\Gamma}^{-1} \widehat{\boldsymbol{\gamma}} \ ,$$

where $\widehat{\boldsymbol{\gamma}} = [\widehat{\gamma}_1 \ \dots \ \widehat{\gamma}_d]'$, $\widehat{\Gamma}$ is the matrix of empirical covariances $\widehat{\Gamma} = (\widehat{\gamma}_{i-j};\ i, j = 1, \dots, d)$, required to be invertible, and $\widehat{\gamma}$ is the empirical covariance function

$$\widehat{\gamma}_\ell = \frac{1}{T} \sum_{t=1}^{T-|\ell|} X_t X_{t+|\ell|} \ .$$

The previous expression supposes that the process is centred, this is $\mathbb{E}[X_t] = 0$ for all t. Consider now $Y = (Y_t)_{t\geq 1}$, an independent copy of X and let $\hat{f}(\boldsymbol{y}_{1:d} | \boldsymbol{X}_{1:T}) = \widehat{\boldsymbol{\theta}}' \boldsymbol{y}_{d:1}$, where $\widehat{\boldsymbol{\theta}}$ is computed from X. Under mild assumptions, including the existence of moments of order 4 for ξ_t (for details see (i)-(iv) in (Bhansali, 1978, Theorem 1)), the following asymptotic result holds (see (Bhansali, 1978, Equation (4.5))) for d and T large enough

$$\mathbb{E}\left[\left(\hat{f}\left(\boldsymbol{Y}_{1:d} | \boldsymbol{X}_{1:T} \right) - Y_{d+1} \right)^2 \right] - \sigma_d^2 \sim M \frac{d}{T} \ , \tag{2.3.6}$$

where $M > 0$ and $\sigma_d^2 = \inf_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}[(\boldsymbol{\theta}' \boldsymbol{Y}_{d:1} - Y_{d+1})^2]$.

Choice of d :

Choosing d arbitrary large comes with the widely known over-fitting issue. Higher dimensional models fit better the training data (also $\sigma_d \searrow \sigma$ when $d \to \infty$) but they are

not suitable for predicting the coming instances of the process. Note that, in particular, d/T (in the right-hand side of (2.3.6)) may become inconveniently large. Several strategies penalising the dimension d have been proposed, such as the Final Prediction Error (FPE) criterion of Akaike (1969), the Akaike Information Criterion (AIC) of Akaike (1973), the bias-corrected version of the AIC (AICC) of Hurvich and Tsai (1989) and the Bayesian Information Criterion (BIC) (see Schwarz (1978) and Akaike (1978)). We also refer to (Brockwell and Davis, 2002, Section 5.5) for an overview.

Example 6 (AR(d)). The real autoregressive process of order d is a particular case of Example 5 (page 41) with $\theta_j = 0$ for all $j > d$, that is

$$X_t = \sum_{j=1}^{d} \theta_j X_{t-j} + \sigma \xi_t \,, \tag{2.3.7}$$

where $\boldsymbol{\theta} = [\theta_1 \ \dots \ \theta_d]' \in \mathbb{R}^d$. Assume moreover that the median of ξ_t is zero.
As in the previous example, consider $Y = (Y_t)_{t \geq 1}$, an independent copy of X. Observe that

$$\sigma_j^2 = \inf_{\boldsymbol{\theta} \in \mathbb{R}^j} \mathbb{E}\left[(\boldsymbol{\theta}' \boldsymbol{Y}_{j:1} - Y_{j+1})^2 \right] = \sigma^2 \quad \text{for} \quad j \geq d \,.$$

The best predictor of the process given its past, with respect to the quadratic loss, is the conditional expectation $\hat{f}(\boldsymbol{y}_{1:t-1}) = \boldsymbol{\theta}' \boldsymbol{y}_{t-1:t-d}$. In general, for any $\sigma(X)$-measurable estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ we have

$$\mathbb{E}\left[\left(\widehat{\boldsymbol{\theta}}' \boldsymbol{Y}_{d:1} - Y_{d+1} \right)^2 \middle| X \right] = \sigma^2 + \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right)' \mathbb{E}\left[\boldsymbol{Y}_{d:1} \boldsymbol{Y}_{d:1}' \right] \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) = \sigma^2 + \| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \|_\Gamma^2 \,, \tag{2.3.8}$$

where $\| \cdot \|_\Gamma$ denotes the norm associated with $\Gamma \in \mathbb{R}^{d \times d}$, the covariance matrix of Y.
This justifies the strategy of looking for efficient estimators of $\boldsymbol{\theta}$ when predicting Y. If we instead consider the ℓ_1 loss, the link between estimation and prediction is less straightforward.

$$\mathbb{E}\left[\left| \hat{f}\left(\boldsymbol{Y}_{1:t-1} \middle| \boldsymbol{X}_{1:T} \right) - Y_t \right| \right] = \mathbb{E}\left[\mathbb{E}\left[\left| \hat{f}\left(\boldsymbol{Y}_{1:t-1} \middle| \boldsymbol{X}_{1:T} \right) - \boldsymbol{\theta}' \boldsymbol{Y}_{t-1:t-d} - \sigma \xi_t \right| \middle| \boldsymbol{X}_{1:T}, \boldsymbol{Y}_{1:t-1} \right] \right] \tag{2.3.9}$$

Observe that ξ_t is independent of $\boldsymbol{X}_{1:T}, \boldsymbol{Y}_{1:t-1}$. The conditional expectation in the right hand-side of Equation (2.3.9) is minimized when $(\hat{f}_t(\boldsymbol{Y}_{1:t-1} | \boldsymbol{X}_{1:T}) - \boldsymbol{\theta}' \boldsymbol{Y}_{t-1:t-d})/\sigma$ is equal to the median value of ξ_t. Since we assume that this median value is zero, the best predictor of the process given its past, with respect to the ℓ_1 loss, is also the conditional expectation $\hat{f}(\boldsymbol{y}_{1:t-1}) = \boldsymbol{\theta}' \boldsymbol{y}_{t-1:t-d}$.
If the $(\xi_t)_{t \in \mathbb{Z}}$ are centred standard Gaussian random variables, the prediction risk satisfies

$$\mathbb{E}\left[\left|\hat{f}_t\left(\boldsymbol{Y}_{1:t-1}|\,\boldsymbol{X}_{1:T}\right)-Y_t\right|\Big|\boldsymbol{X}_{1:T}\right]=\frac{\left(2\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)'\Gamma\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)+2\sigma^2\right)^{1/2}}{\pi^{1/2}}\,, \tag{2.3.10}$$

where $\widehat{\boldsymbol{\theta}}\in\mathbb{R}^d$ is an estimator of $\boldsymbol{\theta}$ that depends only on $\boldsymbol{X}_{1:T}$.

Having available a copy of the process to forecast may be a strong assumption, specially in a dependent context. As mentioned in the beginning of this section, the same approach has been used in practice when X and Y are dependent. The process Y may correspond, for example, to $(X_{T+\Delta+t})_{t\geq 1}$ where Δ is large enough. Another research direction explores the construction of predictors without relying on an independent data set and investigates rigorous bounds on their risks.

2.3.1.2 Predicting a dependent process without an independent data set

Suppose that the observations of the dependent process $X=(X_t)_{t\in\mathbb{Z}}$ arrive one by one. The aim of this section is to present predictors of X_t built exclusively from its past and to provide consistency results under specific conditions.

The learning data set can be the empty set or the past observations $(X_s)_{s\leq 0}$ available before starting the prediction. Let $\mathcal{F}$ be the natural filtration associated with X, that is $\mathcal{F}_t=\sigma(X_s, 1\leq s\leq t)$. We denote by $\hat{f}_t$ the function that predicts X_t from $X_1,\ldots,X_{t-1}$ and $\mathcal{H}$, then, set $\widehat{X}_t=\hat{f}_t(\boldsymbol{X}_{1:t-1}\mid\mathcal{H})=\hat{f}_t((X_s)_{s\leq t-1})$. Let ℓ be the quadratic loss.

If $\mathcal{H}$ is the trivial σ-field, the prediction risk and the risk coincide, and equal

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\hat{f}_t\left(\boldsymbol{X}_{1:t-1}\right)-X_t\right)^2\right]. \tag{2.3.11}$$

Example 7 (Time varying linear processes). Suppose that the parameters $\boldsymbol{\theta}$ and σ vary with t in (2.3.7). Such generalization of Example 6 (page 42) is known as linear processes with time varying coefficients. They are defined by the representation

$$X_t=\sum_{j=1}^{\infty}\theta_j(t)\,X_{t-j}+\sigma(t)\,\xi_t\,, \tag{2.3.12}$$

where (ξ_t) is a sequence of centred independent random variables with variance 1.

As in the previous examples, the estimation of $\boldsymbol{\theta}$ provides the key to the prediction. We need to build $\widehat{\boldsymbol{\theta}}=[\widehat{\theta}_1\ \ldots\ \widehat{\theta}_d]'$, a function from $\mathbb{Z}$ to $\mathbb{R}^d$, with $d\in\mathbb{N}^*$. Stochastic (or online) gradient descent methods became very popular because of their intrinsic

simplicity and proved efficiency. The primary algorithm, adapted to the present example, is sketched in the following.

Algorithm 1: Stochastic gradient descent.

parameters the gradient step size μ;
initialization $t = 0, \widehat{\boldsymbol{\theta}}(t) = [0 \ \ldots \ 0]'$;
while input *a new* X_t*;*
do
  $\widehat{\boldsymbol{\theta}}(t+1) = \widehat{\boldsymbol{\theta}}(t) + \mu \left(X_t - \widehat{\boldsymbol{\theta}}'(t) \boldsymbol{X}_{t-1:t-d} \right)$;
  return $\widehat{\boldsymbol{\theta}}(t+1)$*;*
  $t = t + 1$;

The convergence of the stochastic gradient descent has been widely studied in the stationary case (see Bottou (1998) for an account, and more recently Bottou (2012)). An analysis for individual sequences is provided in Cesa-Bianchi (1999). In contrast, for the class of processes described in this example, proving a prediction risk bound is difficult; the available results are rather sparse.

As explained in Section 2.2.2, the meaningful results available for this kind of models require specific regularity conditions. The TVAR process presented in Example 3 (page 37), where the parameter $\boldsymbol{\theta}$ is β-Hölder continuous with $\beta \in (0, 1]$ is such an example.

In this context, suppose that we have at our disposal enough of observations $(X_s)_{s \leq 0}$. From (Moulines et al., 2005, Theorem 2) we derive the following: let $\widehat{\boldsymbol{\theta}}$ the estimation of $\boldsymbol{\theta}$ obtained from the Normalized Least Square (NLMS) algorithm (a modification of Algorithm 1), there exists a constant $M_1 > 0$ such that

$$\sup_{1 \leq t \leq T} \left(\mathbb{E} \left[\left| \widehat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) \right|^4 \right] \right)^{1/2} \leq M_1 \left(\mu^{1/2} + (T\mu)^{-\beta} \right)^2 .$$

Setting $\hat{f}_t(\boldsymbol{x}_{1:t-1}) = \widehat{\boldsymbol{\theta}}'(t) \boldsymbol{x}_{t-1:t-d}$, or $\widehat{X}_t = \widehat{\boldsymbol{\theta}}'(t) \boldsymbol{X}_{t-1:t-d}$ we conclude that there exists a constant $M_2 > 0$ such that

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\hat{f}_t(\boldsymbol{X}_{1:t-1}) - X_t \right)^2 \right] - \frac{1}{T} \sum_{t=1}^{T} \sigma^2 \left(\frac{t}{T} \right) \leq M_2 \left(\mu^{1/2} + (T\mu)^{-\beta} \right)^2 . \quad (2.3.13)$$

This result is valid for $\beta \in (0, 1]$. A bias reduction technique can be used to obtain the same decaying rate of (2.3.13) for $\beta \in (1, 2]$, see (Moulines et al., 2005, Corollary 9). The step size μ that minimizes the right-hand side of (2.3.13) is proportional to $T^{-2\beta/(2\beta+1)}$. This expression contains β which is usually unknown in practice.

2.3.2 Optimality

In previous sections we describe efficient predictors exhibiting a small prediction risk with high probability or a small risk. In practice, we seek for upper bounds for these risks, and we need them to be as small as possible. What "small" is varies from a situation to another one. Taking up Example 6 (page 37) concerning the AR(d) process, we showed that the prediction risk can not be smaller than σ^2 for the ℓ_2 loss (see Equation (2.3.8)) or smaller than $(2/\pi)^{1/2}\sigma$ for the ℓ_1 loss under the Gaussian assumption (see Equation (2.3.10)). The time has come to make clear and generalize the ideas behind these lower and upper bounds. This section presents the features characterizing the optimality of a forecasting procedure.

Typically, we look for a predictor $\widehat{Z} = (\widehat{Z}_t)_{1\leq t\leq T} \in \mathcal{P}_T$ built from a collection of predictors indexed by Θ: $\widehat{Z}_\Theta = \{\widehat{Z}_{\boldsymbol{\theta}} = (\widehat{Z}_{\boldsymbol{\theta},t})_{1\leq t\leq T} \in \mathcal{P}_T, \boldsymbol{\theta} \in \Theta\}$. It does not necessarily mean that $\widehat{Z} \in \widehat{Z}_\Theta$ because $\widehat{Z}_\Theta$ may be strictly contained in $\mathcal{P}_T$. Let $\widehat{Z}_* = (\widehat{Z}_{*,t})_{1\leq t\leq T}$ be such that

$$\widehat{Z}_* \in \underset{\widetilde{Z}\in\mathcal{P}_T}{\arg\inf}\, R_T\left(\widetilde{Z}\right) ,$$

where the inf is taken over all possible predictors of $(Z_t)_{1\leq t\leq T}$ (see Definition 4). If $\widehat{Z}_*$ corresponds to a $\widehat{Z}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \Theta$, we will say that the model is well-specified. Otherwise it is said to be misspecified.

From our perspective, the comparison with the best of predictors is more informative than the risk itself because it measures how well we can predict exploiting the available knowledge about the process. The following decomposition of the risk raises the recurrent bias-variance tradeoff issue

$$R_T\left(\widehat{Z}\right) - R_T\left(\widehat{Z}_*\right) = \underbrace{\left(R_T\left(\widehat{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right)\right)}_{\text{regret}} + \underbrace{\left(\inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right) - R_T\left(\widehat{Z}_*\right)\right)}_{\text{approximation error}} . \tag{2.3.14}$$

The first term between parenthesis in (2.3.14) corresponds to what we call the regret (or best predictor regret). It measures the pertinence of our choice in Z_Θ. The second term stands for the approximation error and it evaluates the pertinence of the class Z_Θ. To larger classes Z_Θ correspond smaller approximation errors but larger regrets. On the other hand, larger values of T do not impact the approximation error but usually entail smaller regrets. A crucial question is the tradeoff between the size of Z_Θ and T. From a practical point of view, a third error term may appear, which is inherent to the numerical method that we choose to compute $\widehat{Z}$. We can rewrite the regret including $\widetilde{Z}$, the mentioned numerical approximation of $\widehat{Z}$

$$\underbrace{R_T\left(\widetilde{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right)}_{\text{numerical regret}} = \underbrace{\left(R_T\left(\widetilde{Z}\right) - R_T\left(\widehat{Z}\right)\right)}_{\text{numerical approximation error}} + \underbrace{\left(R_T\left(\widehat{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right)\right)}_{\text{regret}} . \tag{2.3.15}$$

In examples 5, 6 and 7 (pages 41, 42 and 43 respectively), having well-specified models is equivalent to assume that the parameter (or the function) $\boldsymbol{\theta}$ generating the process lies on Θ. We have $R_T(\widehat{Z}_*) = \sigma^2$ in examples 5 and 7 and the respective bounds (2.3.6) and (2.3.13) are already expressed as regrets. In Example 6, $R_T(\widehat{Z}_*) = (2/\pi)^{1/2}\sigma$.
In this contribution, the regret is supposed to be the measure of quality of a predictor. The next two sections introduce the oracle inequalities and the minimax approach, both concerning the regret. Then, we present certain MCMC tools that are linked to the numerical approximation error.

2.3.2.1 Oracle inequalities

The notions of *oracle* and *oracle inequalities* were introduced by Donoho and Johnstone (1998). Oracle inequalities provide upper bounds for the regret of a statistical procedure in function of T (see (Tsybakov, 2009, Section 1.8)). These bounds are also, in general, depending on the parameters defining the class of models imposed to Z. Here and in the following, let $\mathcal{M}_{\boldsymbol{\lambda}}$ denote the class of process Z belongs to. It is indexed by the hyperparameter $\boldsymbol{\lambda}$. In Example 5 (page 41) we can consider that $\mathcal{M}_{\boldsymbol{\lambda}}$ is the collection of all processes satisfying Equation (2.3.5) where the generating parameters lie on $s_\infty(\delta)\times\{\sigma\} = \{\{\boldsymbol{\theta} \in \mathbb{R}^{\mathbb{N}^*}, 1 - \sum_{j=1}^{\infty} \theta_j z^j \neq 0, \text{for } |z| \leq \delta^{-1}\} \cap \{|\boldsymbol{\theta}|_1 < \infty\}\} \times \{\sigma\}$, and $\boldsymbol{\lambda} = (\delta, \sigma) \in \mathbb{R}_+^{*2}$.
We specially focus our attention in oracle inequalities holding uniformly on $\mathcal{M}_{\boldsymbol{\lambda}}$, that is when there exist $M_{\boldsymbol{\lambda}} > 0$ only depending on $\boldsymbol{\lambda}$ and a sequence $(\psi_{T,\boldsymbol{\lambda}})_{T\geq 1}$ such that, for all $T \geq 1$

$$\sup_{Z\in\mathcal{M}_{\boldsymbol{\lambda}}} \left\{ R_T\left(\widehat{Z}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\right) \right\} \leq M_{\boldsymbol{\lambda}}\psi_{T,\boldsymbol{\lambda}} \,, \tag{2.3.16}$$

and also when it holds with high probability, this means that there exist $M_{\boldsymbol{\lambda}} > 0$ only depending on $\boldsymbol{\lambda}$ and $(\psi_{T,\boldsymbol{\lambda},\varepsilon})_{T\geq 1}$ such that for all $\varepsilon \in (0,1)$ and $T \geq 1$, with probability at least $1 - \varepsilon$ we have

$$\sup_{Z\in\mathcal{M}_{\boldsymbol{\lambda}}} \left\{ R_T\left(\widehat{Z}\middle|\mathcal{H}\right) - \inf_{\boldsymbol{\theta}\in\Theta} R_T\left(\widehat{Z}_{\boldsymbol{\theta}}\middle|\mathcal{H}\right) \right\} \leq M_{\boldsymbol{\lambda}}\psi_{T,\boldsymbol{\lambda},\varepsilon} \,.$$

The set Θ may be arbitrary. However, there are suitable choices conditioned by the information available about $\boldsymbol{\lambda}$, as for example the set of all $\mathbb{R}^d$ vectors generating a non-explosive autoregressive process in Example 6 (page 42) or a stable Hölder class of functions in Example 7 (page 42). The predictor $\widehat{Z}_{\boldsymbol{\theta}^*}$, where $\boldsymbol{\theta}^* = \arg\inf_{\boldsymbol{\theta}\in\Theta} R_T(\widehat{Z}_{\boldsymbol{\theta}})$, is called the projection oracle, because it gives the best forecast of Z in Θ.

TVAR hyperparameter :

The case of TVAR processes, described in Example 3 (page 37), is of particular interest. Recall that the existing statistical results (including the oracle inequality (2.3.13)) exploit the regularity of $\boldsymbol{\theta}$ (established by β and R) and rely also on its stability (through β, R and δ). The paremeters describing these features, together with the bounds of σ, define $\boldsymbol{\lambda}$.
Let $\beta > 0, \delta \in (0,1), R > 0, \rho \in (0,1]$ and $\sigma_+ > 0$. We say that X belongs to $\mathcal{M}_{\boldsymbol{\lambda}}$ with

$$\lambda = (\beta, R, \delta, \rho, \sigma_+) \ , \tag{2.3.17}$$

if X is a TVAR process generated by a function $\theta \in s_d(\delta) \cap \Lambda_d(\beta, R)$ (see equations (2.2.8) and (2.2.9)), with $\sigma \in [\rho\sigma_+, \sigma_+]$.
The inequality (2.3.13) corresponding to Example 7 holds uniformly for all processes $X \in \mathcal{M}_\lambda$. It fulfills the oracle inequality definition given by (2.3.16).

2.3.2.2 Minimax and adaptiveness

A first and more common question is how fast our method approximates the projection oracle. It is answered via oracle inequalities as explained in Section 2.3.2.1. There is another question that complements the first one: how fast the prediction can be made within $\mathcal{P}_T$? The answer pass through the definition of minimax regret.
The minimax prediction regret is defined by

$$\inf_{\widehat{Z} \in \mathcal{P}_T} \sup_{Z \in \mathcal{M}_\lambda} \left\{ R_T\left(\widehat{Z}\right) - \inf_{\theta \in \Theta} R_T\left(\widehat{Z}_\theta\right) \right\} \ . \tag{2.3.18}$$

No predictor can mimic the projection oracle faster than any lower bound of (2.3.18). Remark that if the model is well-specified, the expression (2.3.18) turns into

$$\inf_{\widehat{Z} \in \mathcal{P}_T} \sup_{Z \in \mathcal{M}_\lambda} \left\{ R_T\left(\widehat{Z}\right) - R_T\left(\widehat{Z}_*\right) \right\} \ .$$

In this situation, all lower bounds are non-negative.
Suppose that there exists a constant $m_\lambda > 0$ only depending on λ and a sequence $(\psi_{T,\lambda})_{T \geq 1}$ such that

$$\inf_{\widehat{Z} \in \mathcal{P}_T} \sup_{Z \in \mathcal{M}_\lambda} \left\{ R_T\left(\widehat{Z}\right) - \inf_{\theta \in \Theta} R_T\left(\widehat{Z}_\theta\right) \right\} \geq m_\lambda \psi_{T,\lambda} \ . \tag{2.3.19}$$

Hence, the fastest we can predict is also lower bounded by this sequence $(\psi_{T,\lambda})_{T \geq 1}$ in (2.3.19). A predictor $\widehat{Z}$ such that the inequality (2.3.16) holds for the very same sequence $(\psi_{T,\lambda})_{T \geq 1}$ is said to be minimax-rate optimal.
There are several investigations about minimax problems in different contexts: nonparametric estimation (see Gill and Levit (1995) and Nemirovskiĭ (1990)), density estimation (see Čencov (1962) and Yang and Barron (1999)), a fixed point for density estimation (see Farrell (1972)). We also point out Birgé (1983) and (Tsybakov, 2009, Chapter 2) and the references therein. The techniques developed beforehand provide an approach to tackle the minimax prediction problem.
The construction of predictors may rely on the knowledge of the parameter λ defining the class $\mathcal{M}_\lambda$, which is usually unknown in practice. In Example 7 (page 43), to

obtain a predictor enjoying the best convergence rate we should, a priori, know β (see Equation (2.3.13) and the remark just below).
More generally, the arsenal of minimax-rate or optimal predictors built when we know λ may be useless when this information is not available. The behavior of a predictor $\widehat{Z}$ may vary from a class $\mathcal{M}_\lambda$ to another one. Methods that circumvent this issue, being minimax-rate for any λ in some set are called adaptive.
Let Λ be a set of possible values of λ. We say that the predictor $\widehat{Z}$ is Λ minimax adaptive if for any $\lambda \in \Lambda$ it is minimax-rate. In other words, for all $\lambda \in \Lambda$, there exist $M_\lambda > 0$ only depending on λ such that for all $T \geq 1$

$$\sup_{Z \in \mathcal{M}_\lambda} \left\{ R_T\left(\widehat{Z}\right) - \inf_{\theta \in \Theta} R_T\left(\widehat{Z}_\theta\right) \right\} \leq M_\lambda \psi_{T,\lambda} \ ,$$

where the sequence $(\psi_{T,\lambda})_{T \geq 1}$ satisfies (2.3.19).
Adaptive methods are suitable because they converge at the same rate as the best of all possible predictors, and without needing very precise information about the process to forecast.
Minimax adaptive procedures date back to the 1980s. Since then, several works investigating different problems have appeared. We refer for example to Efroimovich and Pinsker (1984); Lepskiĭ (1991); Barron and Cover (1991); Donoho and Johnstone (1998); Birgé and Massart (2000); Yang (2000a).

2.3.2.3 Markov chain Monte Carlo methods

In practice, the numerical computation of the predictor $\widehat{Z}$ may lead to a behavior not explained by the bound (2.3.14). If we could exactly calculate $\widehat{Z}$ the question would not arise; otherwise it is useful to investigate how $\widetilde{Z}$, the numerical approximation of $\widehat{Z}$, mimics the projection oracle (see Equation (2.3.15)).
The predictor $\widehat{Z}$ is sometimes given by an integral (see Dalalyan and Tsybakov (2012)). The widely held Markov chain Monte Carlo (MCMC) methods provide a tool-kit to approach it (see Cappé et al. (2005) and Meyn and Tweedie (2009)). Yet, it is crucial to bound the number of iterations that the algorithm needs to achieve a numerical precision of the same order as the prediction risk. A paper of Łatuszyński and Niemiro (2011) contains several results that evaluate the accuracy of a MCMC approximation in function of the number of iterations.
Assume that there exists a function g such that

$$\widehat{Z}_t = \int g(\boldsymbol{u}) \, \pi_0(\mathrm{d}\boldsymbol{u}) \ , \tag{2.3.20}$$

where $\boldsymbol{u}$ belongs to a certain space $\mathcal{U}$ (let us suppose that it is a subspace of $\mathbb{R}^d$ for $d > 0$) endowed with the measure π_0. Consider a Markov chain $U = (U_i)_{i \geq 0}$ with invariant

distribution π_0. Let μ denote the probability distribution of U. We approximate the integral (2.3.20) by

$$\widehat{Z}_{t,n} = \frac{1}{n}\sum_{i=0}^{n-1} g(U_i) \tag{2.3.21}$$

The asymptotic behavior of $\widehat{Z}_{t,n}$ is often investigated via a Central Limit Theorem (CLT) for Markov chains (see Geyer (1992), Jones (2004) and Roberts and Rosenthal (2004)). Asymptotic confidence intervals are established relying on the CLT (we refer to Geyer (1992); Flegal and Jones (2010); Jones and Hobert (2001)).
In contrast, Łatuszyński and Niemiro (2011) proposes an explicit lower bound for n that ensures the following

$$\mu\left(\left|\widehat{Z}_{t,n} - \widehat{Z}_t\right| \leq \alpha\right) \geq 1 - \varepsilon\,,$$

for $\alpha, \varepsilon > 0$, where $\mu(A)$ denotes the probability of A according to the distribution μ.
This lower bound depends on α, ε, the function g and on certain drift condition assumed on the Markov chain U (see (Łatuszyński and Niemiro, 2011, Theorem 3.1)). The drift condition implies (under suitable conditions) the geometric ergodicity (see Meyn and Tweedie (2009) and (Baxendale, 2005, Theorem 1.1)). This is the main ingredient that the proof of (Łatuszyński and Niemiro, 2011, Theorem 3.1) requires.
The more convenient Markov chains U are those that converge faster to the invariant distribution π_0 and provide smaller lower bounds for n. In other words, they allow to approach $\widehat{Z}_t$ at a level α (and specially for $\alpha \propto \psi_{\lambda,T}$ as defined in (2.3.16)) in fewer iterations.

2.4 AGGREGATION

After introducing the models (Section 2.2) and making explicit what we look for with a forecasting procedure (Section 2.3), we present here the approach that we use to propose our predictors.
One of the general machineries for tackling forecasting problems are the aggregation methods. They have been studied for the last 25 years. Aggregation techniques are in the crossroad of the machine learning (see Vovk (1990); Littlestone and Warmuth (1994); Haussler et al. (1998)) and the statistical learning communities (we refer to the seminal works of Barron (1987); Catoni (1997, 2004); Juditsky and Nemirovski (2000); Yang (2000a, 2004); Leung and Barron (2006)). For a recent overview see (Giraud, 2015, Chapter 3).
Popular aggregation algorithms such as Boosting (Freund (1995)), Bagging (Breiman (1996)) and Random Forest (Amit and Geman (1997)) have been widely and successfully applied in practice.

Let Θ be a possibly uncountable set of indexes equipped with a σ-field to be specified and let π be a measure on it called the *prior*. Assume that the observations belong to $\mathcal{X} \subseteq \mathbb{R}$. Assume moreover that $\pi(\Theta) = \int_\Theta \pi(\mathrm{d}\boldsymbol{\theta}) < \infty$. We are provided with a collection $\{(\widehat{x}_t^{(\boldsymbol{\theta})})_{1\leq t\leq T}, \boldsymbol{\theta} \in \Theta\}$, that we call predictors (it is just a terminology, they are not necessarily predictors in the sense of Definition 4). Our first aim is to obtain a new predictor forecasting almost as or more accurately than the best convex combination of $\{(\widehat{x}_t^{(\boldsymbol{\theta})})_{1\leq t\leq T}, \boldsymbol{\theta} \in \Theta\}$ without knowing which is it. A weaker objective is to get a predictor behaving like or better than the best within the provided collection.
Define the simplex

$$\mathcal{S}_\Theta = \left\{ \boldsymbol{s} = (s_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta) \in \mathbb{R}_+^\Theta : \int_\Theta s_{\boldsymbol{\theta}} \pi(\mathrm{d}\boldsymbol{\theta}) = 1 \right\} . \tag{2.4.1}$$

Let $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a loss function. We construct the new predictor from a collection $\{(\alpha_{\boldsymbol{\theta},t})_{1\leq t\leq T}, \boldsymbol{\theta} \in \Theta\}$ such that for any $1 \leq t \leq T$, $\boldsymbol{\alpha}_t = (\alpha_{\boldsymbol{\theta},t}, \boldsymbol{\theta} \in \Theta) \in \mathcal{S}_\Theta$. We present formally our two objectives.
The first one is to build $\{(\alpha_{\boldsymbol{\theta},t})_{1\leq t\leq T}, \boldsymbol{\theta} \in \Theta\}$ such that

$$\frac{1}{T}\sum_{t=1}^T \ell\left(\int_\Theta \alpha_{\boldsymbol{\theta},t}\widehat{x}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta}), x_t\right) - \inf_{\nu\in\mathcal{S}_\Theta} \frac{1}{T}\sum_{t=1}^T \ell\left(\int_\Theta \nu_{\boldsymbol{\theta}}\widehat{x}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta}), x_t\right) , \tag{2.4.2}$$

is as small as possible (known as the convex regret bounds problem).
The second objective is to propose $\{(\alpha_{\boldsymbol{\theta},t})_{1\leq t\leq T}, \boldsymbol{\theta} \in \Theta\}$ such that

$$\frac{1}{T}\sum_{t=1}^T \ell\left(\int_\Theta \alpha_{\boldsymbol{\theta},t}\widehat{x}_t^{(\boldsymbol{\theta})}\pi(\mathrm{d}\boldsymbol{\theta}), x_t\right) - \inf_{\boldsymbol{\theta}\in\Theta} \frac{1}{T}\sum_{t=1}^T \ell\left(\widehat{x}_t^{(\boldsymbol{\theta})}, x_t\right) , \tag{2.4.3}$$

is as small as possible (known as the best predictor regret bounds problem).
Expressions (2.4.2) and (2.4.3) recall the regret introduced in (2.3.14). Their stochastic versions are detailed in Section 2.4.2. When ℓ is the quadratic loss, two exponentially weighted aggregation strategies are extensively studied.
Strategy 1: building weights from the gradient of the quadratic loss : The first strategy consists in defining for all $\boldsymbol{\theta} \in \Theta$ and $t = 1, \ldots, T$, the weights $\widehat{\alpha}_{\boldsymbol{\theta},t}$ by

$$\widehat{\alpha}_{\boldsymbol{\theta},t} = \frac{\exp\left(-2\eta\sum_{s=1}^{t-1}\left(\int_\Theta \widehat{\alpha}_{\boldsymbol{\theta}_1,s}\widehat{x}_s^{(\boldsymbol{\theta}_1)}\pi(\mathrm{d}\boldsymbol{\theta}_1) - x_s\right)\widehat{x}_s^{(\boldsymbol{\theta})}\right)}{\int_\Theta \exp\left(-2\eta\sum_{s=1}^{t-1}\left(\int_\Theta \widehat{\alpha}_{\boldsymbol{\theta}_1,s}\widehat{x}_s^{(\boldsymbol{\theta}_1)}\pi(\mathrm{d}\boldsymbol{\theta}_1) - x_s\right)\widehat{x}_s^{(\boldsymbol{\theta}_2)}\right)\pi(\mathrm{d}\boldsymbol{\theta}_2)} , \tag{2.4.4}$$

with the convention that a sum over no element is null, so $\widehat{\alpha}_{\boldsymbol{\theta},1} = 1/\pi(\Theta)$ for all $\boldsymbol{\theta}$.
The parameter $\eta > 0$, is usually called the *learning rate*. We set its value in function of the specific framework on the observations. This strategy provides guarantees for

the aggregated predictor compared to the best of the constant convex combinations of predictors. Its downside is that the regret is of the order of $T^{-1/2}$ (the details can be found in the proof of inequality (4.2.12) of Lemma 5, page 91).

Strategy 2: building weights from the quadratic loss : The second strategy consists in defining for all $\theta \in \Theta$ and $t = 1, \dots, T$, the weights $\widehat{\alpha}_{\theta,t}$ by

$$\widehat{\alpha}_{\theta,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(\theta)} - x_s\right)^2\right)}{\int_\Theta \exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(\theta_1)} - x_s\right)^2\right) \pi(\mathrm{d}\theta_1)}, \tag{2.4.5}$$

with again the convention that a sum over no element is null. When observations and predictors are bounded for example, this strategy exhibits a regret decaying as T^{-1} for a well-chosen η (see inequality (4.2.13) in Lemma 5, page 91). The result comes from the exp-concavity of the quadratic loss (we refer to (Cesa-Bianchi and Lugosi, 2006, Section 3.3) and (Catoni, 2004, Proposition 2.2.1)). It is closely related to several developments in the stochastic framework; see, for instance, Example 10. The regret in this case is computed with respect to the best predictor. From this point of view, the result is weaker than that obtained using the weights (2.4.4).

2.4.1 Sequential prediction

In contrast to the statistical viewpoint, the theory of individual sequences does not assume that the observations are the realization of a stochastic process (see Cesa-Bianchi and Lugosi (2006)). From this perspective, the online learning (also called sequential prediction) provides an algorithmic tool-kit for addressing problems in statistical learning. We refer to the works of Foster (1991); Auer et al. (2002); Vovk (2006); Stoltz (2011); Gerchinovitz (2013) on online regression for arbitrary sequences.

Example 8. Suppose that $\Theta = \{1, \dots, N\}$ and that $\pi(k) = 1$ for all $k = 1, \dots, N$. Equation (2.4.5) turns into

$$\widehat{\alpha}_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(i)} - x_s\right)^2\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(k)} - x_s\right)^2\right)}. \tag{2.4.6}$$

Assume that observations and predictions belong to the interval $[-B, B]$ with $B > 0$. In this context, (Cesa-Bianchi and Lugosi, 2006, Theorem 3.2 and Proposition 3.1) ensure that for all $0 < \eta < 1/(8B^2)$ and $T > 0$

$$\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{N} \widehat{\alpha}_{i,t} \widehat{x}_t^{(i)} - x_t\right)^2 - \min_{1 \le i \le N} \frac{1}{T} \sum_{t=1}^{T} \left(\widehat{x}_t^{(i)} - x_t\right)^2 \le \frac{\log N}{T\eta}, \tag{2.4.7}$$

where $\widehat{\alpha}_{i,t}$ is defined by Equation (2.4.6). We refer to Inequality (4.2.13) in Lemma 5 for a generalization.

The sequential and the stochastic prediction contexts not only share techniques. The first one allows also to understand which terms of the obtained guaranties for the second one (oracle inequalities) are explained by the statistical assumptions and which terms are inherent to the procedure (the aggregation in our case).

2.4.2 Stochastic prediction

As explained in Section 2.3.2.1, when we impose a stochastic model on the observations, we pose the regret bound problem in terms of expectation and conditional expectation. In this section we present interesting results obtained in this setting.
Provided $\{\widehat{X}^{(\theta)}, \theta \in \Theta\}$, with $\widehat{X}^{\theta} = (\widehat{X}_t^{(\theta)})_{1\le t\le T}$, and $\alpha = \{(\alpha_{\theta,t})_{1\le t\le T}, \theta \in \Theta\}$, with $\alpha_t = (\alpha_{\theta,t}, \theta \in \Theta) \in \mathcal{S}_\Theta$, we let $\widehat{X}^{[\alpha]} = (\widehat{X}_t^{[\alpha_t]})_{1\le t\le T}$ denote the aggregated predictor defined as

$$\widehat{X}_t^{[\alpha_t]} = \int_\Theta \alpha_{\theta,t}\widehat{X}_t^{(\theta)}\pi(\mathrm{d}\theta) \ . \tag{2.4.8}$$

For a $\nu \in \mathcal{S}_\Theta$ we use the same notation $\widehat{X}^{[\nu]} = (\widehat{X}_t^{[\nu]})_{1\le t\le T}$, where

$$\widehat{X}_t^{[\nu]} = \int_\Theta \nu_\theta\widehat{X}_t^{(\theta)}\pi(\mathrm{d}\theta) \ . \tag{2.4.9}$$

Observe that in contrast to (2.4.8), in the expression (2.4.9) the weight ν_θ is the same for all $t = 1, \dots, T$.
The convex regret bounds problem seeks to upper bound

$$R_T\left(\widehat{X}^{[\alpha]}\right) - \inf_{\nu\in\mathcal{S}_\Theta} R_T\left(\widehat{X}^{[\nu]}\right) \ .$$

In the case of the best predictor regret bounds, the expression to be upper bounded is

$$R_T\left(\widehat{X}^{[\alpha]}\right) - \inf_{\theta\in\Theta} R_T\left(\widehat{X}^{(\theta)}\right) \ .$$

For the sake of brevity, we do not write the corresponding conditional regret bounds.
The weights $\alpha_{\theta,t}\pi(\mathrm{d}\theta)$ may depend on $(X_s)_{s<t}, \{\widehat{X}_s^{(\theta)}, s < t, \theta \in \Theta\}$ and also on the learning σ-field $\mathcal{H}$ (possibly independent of the observations). This extra randomness put on the weights has a PAC-Bayesian flavor.
The Probably Approximately Correct (PAC) learning framework, introduced by Valiant (1984), provides guarantees on the approximation error of a statistic that hold with high probability regarding the representativeness of the learning set (in our framework, we could interpret it as $\mathcal{H}$). The Bayesian statistical modeling relies on the prior distribution

that we impose to the unknown parameters. PAC-Bayesian inequalities are inspired in these two theories and were introduced by McAllester (1999). A few years later, Audibert (2004), Catoni (2004) and Dalalyan and Tsybakov (2008) proved PAC-Bayesian inequalities on aggregated statistical procedures.

Example 9. Consider the classical setting where we are given with an i.i.d. sample $((X_i, Y_i))_{1\leq i\leq n}$ and we want to predict the incoming Y_{n+1} in function of X_{n+1} and the learning set. Suppose that we count on a set of predictors $\widehat{Y}^{(k)} : \mathcal{X} \to \mathcal{Y}$ indexed by $k \in \Theta = \{1, \dots, N\}$, and that we set $\pi(k) = 1$ for all k. Assume moreover that there exists $B > 0$ bounding almost surely Y and the predictions $\widehat{Y}^{(k)}$ for all k.
There exists $M > 0$ such that the predictor $\widehat{Y}^{[\alpha]}$ constructed from the exponential weights α computed in (Audibert, 2004, Section 4.2.2, Chapter 1) satisfies for any $\varepsilon > 0$

$$R\left(\widehat{Y}^{[\alpha]}\middle|\mathcal{H}\right) - \inf_{\nu\in\mathcal{S}_\Theta} R\left(\widehat{X}^{[\nu]}\middle|\mathcal{H}\right) \leq M\left(\frac{\log\left(N\log\left(2n\right)/\varepsilon\right)}{n}\right)^{1/2} + M\left(\frac{\log\left(N\log\left(2n\right)/\varepsilon\right)}{n}\right). \quad (2.4.10)$$

Example 10. Suppose now that we observe $((X_t, Y_t))_{t\geq 1}$, instances of a possible dependent and non-stationary process. The context is similar to that described in Section 2.3.1.2. Consider the following decomposition

$$Y_t = \mathbb{E}\left[Y_t \middle| X_t, (X_s, Y_s)_{s\leq t-1}\right] + \xi_t \,. \quad (2.4.11)$$

Assume that we are given with a countable collection of predictors $\widehat{Y}^{(k)} : \mathcal{X} \to \mathcal{Y}$ indexed by $k \in \Theta = \mathbb{N}^*$ and at a bounded distance from the conditional mean (see the right-hand side of Equation (2.4.11)). Under an exponential moment condition on the noise ξ, (Yang, 2004, Theorem 5) ensures that for $\eta > 0$ small enough, the predictor $\widehat{Y}^{[\widehat{\alpha}]}$, built from the weights $\widehat{\alpha}$ defined as in (2.4.5), satisfies

$$R\left(\widehat{Y}^{[\widehat{\alpha}]}\right) \leq \inf_{k\geq 1}\left\{\frac{\log\left(1/\pi_k\right)}{\eta T} + R\left(\widehat{Y}^{(k)}\right)\right\},$$

where $\sum_{k=1}^{\infty}\pi_k = 1$.

The optimality of the remaining term of the aggregation (e.g. the right-hand side of (2.4.10)) have been investigated, specially in the i.i.d. setting. We present a well known lower bound on the remaining term in the context of the estimation of a regression function.

Example 11. Consider the regression model

$$Y_i = f(X_i) + \xi_i \ ,$$

where the $(X_i)_{1\leq i\leq n}$ are i.i.d. random vectors and $(\xi_i)_{1\leq i\leq n}$ are real i.i.d. centred Gaussian, independent of $(X_i)_{1\leq i\leq n}$. Let $\mathcal{F}_0 = \{f : |f|_\infty \leq L\}$, for $L > 0$ and $\Theta = \{1, \dots, N\}$. Under mild assumptions (Tsybakov, 2003, Theorem 2) guaranties that there exists $c > 0$ such that

$$\sup_{f_1,\dots,f_N\in\mathcal{F}_0} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}_0} \left\{ \mathbb{E}\left[\left(\hat{f}_n(X) - f(X)\right)^2\right] - \min_{\theta\in S_\Theta} E\left[\left(\sum_{k=1}^{N} \theta_k f_k(X) - f(X)\right)^2\right]\right\} \geq c\psi_n \ ,$$

where the inf is taken over all the estimators, that is the $\sigma((X_i, Y_i)_{1\leq i\leq n})$-measurable real functions $\hat{f}_n$, X is supposed to be independent of $\sigma((X_i, Y_i)_{1\leq i\leq n})$ and

$$\psi_n = \begin{cases} N/n & \text{, if } N \leq n^{1/2} \ , \\ \left((1/n)\log\left(1 + N/n^{1/2}\right)\right)^{1/2} & \text{, if } N > n^{1/2} \ . \end{cases}$$

To go deeper into it, we also refer to the contributions of Juditsky and Nemirovski (2000); Yang (2004); Audibert (2009).

2.5 QUESTIONS OF THE THESIS

Chapter 3 deals with Causal Bernoulli Shifts. Our main question is how to predict a CBS $Y = (Y_t)_{t\geq 1}$ given a learning data set $X = (X_t)_{1\leq t\leq T}$ and a possibly infinite collection of predictors $\{f_\theta, \theta \in \Theta\}$ as presented in sections 2.3.1.1 and 2.3.2. In a nutshell, we provide a PAC-Bayesian oracle inequality of the prediction risk and a PAC-Bayesian oracle inequality that applies to the numerical computation of the aggregated predictor.
In Chapter 4 we introduce the sub-linear processes, which are, in general, dependent, non-stationary and not uniformly bounded. Considering a finite number of predictors, and without a learning data set (as in Section 2.3.1.2) we investigate how to forecast this kind of process relying on the aggregation.
In the particular case of locally stationary time varying autoregressive (TVAR) processes (see Example 3, page 37) we look for minimax adaptive predictors. Let $R, \delta, \rho, \sigma_+ > 0$, and let J be a compact subset of $\mathbb{R}_+^*$. We precisely seek for Λ minimax adaptive predictors (see Section 2.3.2.2) where $\Lambda = \{(\beta, R, \delta, \rho, \sigma_+) : \beta \in J\}$.
As brought forward in Example 7 (page 43), minimax-rate predictors of a TVAR process are available when the regularity β belongs to $(0, 2]$. Chapter 5 proposes in particular new minimax-rate predictors for $\beta \geq 2$. The aim of the chapter is more general, we investigate

the regression problem in a locally stationary context.

2.6 Main results

2.6.1 Causal Bernoulli Shifts

In Chapter 3 we suppose that we have observed a CBS $(X_t)_{1\leq t\leq T}$ distributed as P and that we wish to predict an independent copy of it, say $(Y_t)_{1\leq t\leq T}$. At each moment $1 \leq t \leq T$ we have access to all the observations $(X_t)_{1\leq t\leq T}$ and to a certain set of predictions $f_\Theta = \{f_\theta, \theta \in \Theta\}$. For making their forecasts at moment t, the predictors in f_Θ may have access to the previous samples $(Y_s)_{s<t}$ but we do not. It is an interesting feature of our framework: we do not exploit directly the sequence $(Y_t)_{1\leq t\leq T}$ but only through f_Θ. For the sake of simplicity we assume that the sequences are real valued.
Let ℓ denote the loss function. The set Θ is also indexed by, and possibly changing with T. Let $d_T > 0$ and suppose that for all $\theta \in \Theta_T$ the function f_θ is defined on $\mathbb{R}^{d_T}$, this means that for any $t > d_T$ the prediction of Y_t that corresponds to θ is given by $f_\theta(\boldsymbol{Y}_{t-1:t-d_T})$. We endow Θ_T with the prior measure π_T and build the Gibbs predictors

$$\hat{f}_{\eta,T}(\cdot|X) = \int_\Theta \nu_\theta(\eta, T, X)\, f_\theta(\cdot)\, \pi_T(\mathrm{d}\theta) \ , \tag{2.6.1}$$

where the aggregation weights depend only on the learning rate η, the learning data set X and its size T. The coefficient ν_θ satisfies the following (see Alquier and Wintenberger (2012))

$$\nu_\theta(\eta, T, X) \propto \exp\left(-\frac{\eta}{T-d_T}\sum_{t=d_T+1}^{T} \ell\left(f_\theta(X_{t-1:t-d_T}), X_t\right)\right) , \tag{2.6.2}$$

and

$$\int_\Theta \nu_\theta(\eta, T, X)\, \pi_T(\mathrm{d}\theta) = 1 \ . \tag{2.6.3}$$

Under mild assumptions on the innovations generating the process X, the collection of predictors, the loss function ℓ, the set Θ_T and the measure π_T, the following PAC-Bayesian oracle inequality holds for all $\varepsilon \in (0,1)$ with P-probability at least $1-\varepsilon$

$$R\left(\hat{f}_{\eta_T,T}(\cdot|X)\right) \leq \inf_{\theta\in\Theta_T} R(f_\theta) + \mathcal{E}\frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}}\log\left(\frac{1}{\varepsilon}\right) , \tag{2.6.4}$$

where $\eta_T = \log T$ and the constant $\mathcal{E}$ is explicitly computed from the assumptions.

This inequality applies to the exact aggregated predictor $\hat{f}_{\eta_T,T}(\cdot|X)$. In practice, the integral (2.6.1) is numerically approximated by $\bar{f}_{\eta,T,n}(\cdot|X) = \sum_{i=0}^{n-1} f_{\theta_i}/n$ where $(\boldsymbol{\theta}_i)_{i\geq 0}$ are the instances of a Markov chain $\Phi_{\eta,T}(X)$ having $\nu_\cdot(\eta, T, X)\pi_T$ as unique invariant measure. This Markov chain is typically constructed using a MCMC method. The Metropolis-Hastings algorithm is such an example.

A Markov chain adds a second source of randomness to the forecasting process. We define $\nu_{\eta,T}$, a probability distribution on $(X, \Phi_{\eta,T}(X))$. Supposing that $\Phi_{\eta,T}(X)$ is geometrically ergodic and under the assumptions that lead to inequality (2.6.4), we prove that for all $\varepsilon \in (0,1)$ and $n \geq M(T,\varepsilon)$, with $\nu_{\eta_T,T}$-probability at least $1-\varepsilon$ we have

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot|X)\right) \leq \inf_{\theta\in\Theta_T} R(f_\theta) + \left(\mathcal{E} + \frac{2}{\log 2} + 2\right)\frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}}\log\left(\frac{1}{\varepsilon}\right), \quad (2.6.5)$$

where $\eta_T = \log T$, $\mathcal{E}$ is the same of inequality (2.6.4) and $M(T,\varepsilon)$ depends in particular, on the convergence rate of $\Phi_{\eta_T,T}(X)$ to its invariant distribution.

Observe that the right-hand sides of inequalities (2.6.4) and (2.6.5) are of the same order. To the best of our knowledge, bounds like (2.6.5) have not been studied before for aggregation procedures in a PAC-Bayesian context when Θ is potentially not finite.

To illustrate our result, we consider the simple case of a real valued stable autoregressive process of finite order d (as in Example 6, page 42) with unit normally distributed innovations. Let $\ell(x,y) = |x-y|$, $d_T = \lfloor \log T \rfloor$, $\Theta_T \subset \mathbb{R}^{d_T}$ and $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}'\boldsymbol{x}$ for any $\boldsymbol{x} \in \mathbb{R}^{d_T}$. For the precise definition of Θ_T and the prior π_T we refer to Section 3.5.

The aggregated predictor is also linear and can be expressed as $\hat{f}_{\eta,T}(\boldsymbol{x}|X) = \widehat{\boldsymbol{\theta}}'_{\eta,T}(X)\boldsymbol{x}$, where

$$\widehat{\boldsymbol{\theta}}_{\eta,T}(X) = \int_\Theta \nu_{\boldsymbol{\theta}}(\eta, T, X)\,\boldsymbol{\theta}\,\pi_T(\mathrm{d}\boldsymbol{\theta}) \ ,$$

with ν defined as in (2.6.2)-(2.6.3). We use the Metropolis-Hastings algorithm to approximate $\widehat{\boldsymbol{\theta}}_{\eta,T}(X)$ (the details are given in Section 3.5) by $\bar{\boldsymbol{\theta}}_{\eta,T,n}$. Let γ_0 be the variance of the process X. For a number of iterations n bigger than $M^*(T,\varepsilon) = 9\gamma_0^3 T^2 \exp(\gamma_0 T/16)/(2\pi\varepsilon^2\log^3 T)$ we guaranty that the bound (2.6.5) is reached. This possibly pessimistic upper bound for $M(T,\varepsilon)$ makes the procedure computationally prohibitive. The predictor $\bar{\boldsymbol{\theta}}'_{\eta_T,T,n}\boldsymbol{x}$ ($T = 2^{12}$ and $n = 1000$) exhibits a poor behavior in our numerical experiences.

2.6.2 Non stationary sub-linear processes and time varying autoregressive processes

Chapter 4 provides general results on sub-linear processes. We go deeper in the study of the particular case of TVAR processes.

Aggregation bounds for sub-linear processes :

Consider a real valued sub-linear sequence $X = (X_t)_{t\in\mathbb{Z}}$ with respect to the noise $(Z_t)_{t\in\mathbb{Z}}$. Recall that

$$|X_t| \leq \sum_{j\in\mathbb{Z}} A_t(j)\, Z_{t-j} \ ,$$

where $(A_t(j))_{t,j\in\mathbb{Z}}$ are non-negative coefficients satisfying

$$A_* := \sup_{t\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} A_t(j) < \infty \ .$$

Suppose that the X_ts arrive on the fly. At each moment $1 \leq t \leq T$ we have access to $(X_s)_{1\leq s\leq t-1}$ and to a certain set of predictions $\{X_t^{(i)}, i = 1, \ldots, N\}$ and we want to build our own online predictor of X_t.

Relying on purely deterministic oracle inequalities derived from (Stoltz, 2011, Theorem 1.7) and (Catoni, 2004, Proposition 2.2.1), a uniform bound on the ℓ_1 norm of the time varying sub-linear coefficients, a Lipschitz assumption on the predictors and moment conditions on the noise appearing in the linear representation of X, we obtain the following oracle inequalities.

(i) Consider a noise Z with finite 4th-order moment and let $\widehat{X} = (\widehat{X}_t)_{1\leq t\leq T}$ denote the aggregated predictor obtained using the weights (2.4.4) with $\eta \propto ((\log N)/T)^{1/2}$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \inf_{\nu\in\mathcal{S}_N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + C_1 \left(\frac{\log N}{T}\right)^{1/2} , \qquad (2.6.6)$$

with the constant C_1 that can be computed from the assumptions.

(ii) Assume that the noise Z has a finite pth-order moment for a given $p > 2$ and let $\widehat{X} = (\widehat{X}_t)_{1\leq t\leq T}$ denote the aggregated predictor obtained using the weights (2.4.5) and $\eta \propto ((\log N)/T)^{2/p}$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \min_{1\leq i\leq N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + C_2 \left(\frac{\log N}{T}\right)^{1-2/p} , \qquad (2.6.7)$$

where C_2 is explicitly computed from the assumptions.

(iii) Suppose that the noise Z has a finite exponential moment and let $\widehat{X} = (\widehat{X}_t)_{1\leq t\leq T}$ denote the aggregated predictor obtained using the weights (2.4.5) with $\eta \propto (\log(T/(\log N)))^{-2}$. Then, when $(\log N)/T \to 0$ we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \min_{1\leq i\leq N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + C_3 \frac{\log N}{T}\left(\log\left(\frac{T}{\log N}\right)\right)^2 , \qquad (2.6.8)$$

and the constant C_3 depends on the assumptions.

Yang (2004) proposed best predictor regret bounds in the context of sequences of possibly dependent random variables. One of the main ingredients of that paper is that the predictors are assumed to remain at a bounded distance to the conditional means. Inequality (2.6.8) is comparable to one of its results, but we obtain it under milder assumptions.

Even though the i.i.d. setting and ours are dissimilar, we present a short comparison of results in both frameworks. Concerning the convex regret bound, Example 11 (Tsybakov (2003)) provides the best possible remaining term when the predictors are bounded. It is (roughly) $(\log N/T)^{1/2}$ if N is much larger than $T^{1/2}$ and N/T when N is smaller than $T^{1/2}$. Hence our bound (2.6.6) coincides only in the case where N is much larger than $T^{1/2}$. However, when N is smaller than $T^{1/2}$, a more complex aggregation procedure allows to get a convex regret bound with a remaining term of the order of $N(\log T)^3/T$ (see (4.9.7) page 125) if the noise has a finite exponential moment. On the other hand, imposing moment conditions of order p on the noise and relying on a uniform bound on the predictors, Audibert (2009) shows that the optimal aggregation rate is $(\log N/T)^{1-2/(p+2)}$ which is slightly smaller than our $(\log N/T)^{1-2/p}$ in (2.6.7).

Aggregation bounds for TVAR processes :

In the context of TVAR processes (see Example 3, page 37), let $\beta > 0$, $\delta \in (0,1)$, $R > 0$, $\rho \in (0,1]$ and $\sigma_+ > 0$ define the hyperparameter $\boldsymbol{\lambda} = (\beta, R, \delta, \rho, \sigma_+)$ indexing the class $\mathcal{M}_{\boldsymbol{\lambda}}$. In Section 4.3.2 we provide a lower bound for the minimax prediction risk (2.3.19). For T large enough we obtain that

$$\inf_{\widehat{X}\in\mathcal{P}_T} \sup_{X\in\mathcal{M}_{\boldsymbol{\lambda}}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] - \frac{1}{T}\sum_{t=1}^{T} \sigma^2\left(\frac{t}{T}\right) \right\} \geq m_{\boldsymbol{\lambda}} T^{-2\beta/(2\beta+1)} . \tag{2.6.9}$$

This rate coincides with that exhibited by the predictor build from the NLMS estimator and a well chosen gradient step size if $\beta \in (0,2]$. Hence, $T^{-2\beta/(2\beta+1)}$ is the optimal minimax rate for $\mathcal{M}_{\boldsymbol{\lambda}}$ processes (at least if $\beta \in (0,2]$).

Let $\beta_0 \in (0,\infty]$ and let $\{\widehat{X}^{(\beta)}, \beta \in (0,\beta_0)\}$ be a collection of β-minimax-rate predictors (δ, R, ρ and σ_+ being fixed). If $\beta_0 < \infty$ we set $N = \lceil \log T \rceil$ and select $\beta_i = (i-1)\beta_0/N$ for $i = 1,\dots,N$. Otherwise we set $N = \lceil (\log T)^2 \rceil$ and $\beta_i = (i-1)\beta_0/N^{1/2}$ for $i = 1,\dots,N$. To construct $\widehat{X}$, we aggregate the predictors $\{\widehat{X}^{(\beta_i)}, i = 1,\dots,N\}$, each of them enjoying the optimal minimax convergence rate for their respective superscript β, using the weights (2.4.5) and the learning rate chosen as follows

(i) if the noise Z has a finite pth-order moment for a given $p > 2$ and $\beta_0 \leq (p-2)/4$, let $\eta \propto (\log(\lceil \log T \rceil)/T)^{2/p}$,

(ii) if the noise Z has a finite exponential moment, let $\eta \propto (\log T)^{-3}$.

Let $\Lambda = \{(\beta, R, \delta, \rho, \sigma_+) : \beta \in (0,\beta_0)\}$. We show, with the help of the oracle inequalities enunciated before, that $\widehat{X} = (\widehat{X}_{t,T})_{1\leq t\leq T}$ is Λ minimax adaptive, this means that

$$\sup_{X\in\mathcal{M}_\lambda}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T}-X_{t,T}\right)^2\right]-\frac{1}{T}\sum_{t=1}^{T}\sigma^2\left(\frac{t}{T}\right)\right\}\leq M_\lambda T^{-2\beta/(2\beta+1)}\ ,$$

for any $\boldsymbol{\lambda}\in\Lambda$.

An important characteristic of $\widehat{X}$ is that it can be calculated recursively and is thus applicable in an online prediction context. The following algorithm details the procedure using the NLMS, supposing that $\beta_0=1$ and that the noise has a finite exponential moment.

Algorithm 2: Online adaptive prediction

parameters the value of T, the order d;
initialization $X_{s,T}=0$ for $-d\leq s\leq 0$, $\eta=(\log T)^{-3}$, $N=\lceil\log T\rceil$, $\widehat{\boldsymbol{\theta}}_{i,-1,T}=\mathbf{0}\in\mathbb{R}^d$ for $i=1,\dots,N$, $t=1$, $\widehat{\alpha}_t=(1/N)_{i=1,\dots,N}$;
while input $X_{t-1,T}$ *is provided;*
do
  for $i=1$ **to** N **do**
    $\beta_i=(i-1)/N$;
    $\mu_i=T^{-2\beta_i/(2\beta_i+1)}$;
    $\widehat{\boldsymbol{\theta}}_{i,t-1,T}=\widehat{\boldsymbol{\theta}}_{i,t-2,T}+\mu_i\left(X_{t-1,T}-\widehat{\boldsymbol{\theta}}'_{i,t-2,T}\boldsymbol{X}_{t-2:t-d-1,T}\right)\dfrac{\boldsymbol{X}_{t-2:t-d-1,T}}{1+\mu_i\left|\boldsymbol{X}_{t-2:t-d-1,T}\right|^2}$;
    for $k=1$ **to** d **do**
      $\widehat{\boldsymbol{\theta}}_{i,t-1,T}(k)=\min\left\{\max\left\{-\binom{n}{k},\widehat{\boldsymbol{\theta}}_{i,t-1,T}(k)\right\},\binom{n}{k}\right\}$;
    $\widehat{X}^{(i)}_{t,T}=\widehat{\boldsymbol{\theta}}'_{i,t-1,T}\boldsymbol{X}_{t-1:t-d,T}$;
  $\widehat{X}_{t,T}=\widehat{X}^{\lfloor\widehat{\alpha}_t\rfloor}_{t,T}=\sum_{i=1}^{N}\widehat{\alpha}_{i,t}\widehat{X}^{(i)}_{t,T}$;
  return $\widehat{X}_{t,T}$;
  $t=t+1$;
  while input *the value of* $X_{t-1,T}$*;*
  do
    for $i=1$ **to** N **do**
      $v_{i,t}=\widehat{\alpha}_{i,t-1}\exp\left(-\eta\left(\widehat{X}^{(i)}_{t-1,T}-X_{t-1,T}\right)^2\right)$;
    $\widehat{\alpha}_t=\left(v_{i,t}/\sum_{k=1}^{N}v_{k,t}\right)_{i=1,\dots,N}$;

2.6.3 Locally stationary processes

Let $d\in\mathbb{N}^*,\beta\geq 2,R,f_->0$ and $\boldsymbol{\lambda}=(\beta,R,f_-)$. In Chapter 5 our study concerns $\Lambda'_1(\beta,R)$, a subset of $\Lambda_1(\beta,R)$ (see Section 5.2.2). Consider $\mathcal{M}_\lambda$ the set of all locally stationary processes according to Definition 14 (it generalizes those characterized by Equation (5.2.4)), such that their local spectral densities $f(\cdot,\omega)\in\Lambda'_1(\beta,R)$ for all ω and $f\geq f_-$ (see Definition 13). We address the following regression problem:

$$\boldsymbol{\theta}^*_{t,T} = \underset{\boldsymbol{\theta}=[\theta_1 \dots \theta_d]\in\mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[\left(X_{t,T} - \sum_{k=1}^{d} \theta_k\, X_{t-k,T}\right)^2\right] = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}\left[\left(X_{t,T} - \boldsymbol{\theta}' \boldsymbol{X}_{t-1:t-d,T}\right)^2\right] ,$$

where $X \in \mathcal{M}_\lambda$ and B' denotes the transpose of matrix B.
The vector $\boldsymbol{\theta}^*_{t,T}$ coincides with $\boldsymbol{\theta}(t/T)$ in the case of locally stationary TVAR processes.
Given $h : [0,1] \to \mathbb{R}$ and $m \le T$, the empirical local covariance function $\widehat{\gamma}_m$ is defined in $\mathbb{R} \times \mathbb{Z}$ as

$$\widehat{\gamma}_m(u,\ell) = \frac{1}{H_m} \sum_{\substack{t_1,t_2=1 \\ t_1-t_2=\ell}}^{m} h\left(\frac{t_1}{m}\right) h\left(\frac{t_2}{m}\right) X_{\lfloor uT\rfloor+t_1-m/2,T} X_{\lfloor uT\rfloor+t_2-m/2,T} , \tag{2.6.10}$$

where $H_m = \sum_{k=1}^{m} h^2(k/m)$.
Relying on the previous definition we propose an estimator $\widetilde{\boldsymbol{\theta}}_{t,T}$ of $\boldsymbol{\theta}^*_{t,T}$ in two steps. Let $k = \lceil \beta \rceil - 1$ and $M \in 2^{k+1}\mathbb{N}^*$. In the first step we use the Yule-Walker equations (see Dahlhaus and Giraitis (1998)) and for $m \in \{M/2^j, j = 0, \dots, k\}$ we construct

$$\widehat{\boldsymbol{\theta}}_{t,T}(m) \ = \ \widehat{\Gamma}^{-1}_{t,T,m} \widehat{\boldsymbol{\gamma}}_{t,T,m} ,$$

where $\widehat{\boldsymbol{\gamma}}_{t,T,m} = [\widehat{\gamma}_m(t/T,1) \ \dots \ \widehat{\gamma}_m(t/T,d)]'$, $\widehat{\Gamma}_{t,T,m}$ is the matrix of empirical covariances $\widehat{\Gamma}_{t,T,m} = (\widehat{\gamma}_m(t/T, i-j);\, i,j = 1,\dots,k)$ and $\widehat{\gamma}_m$ is the empirical covariance function as defined in (2.6.10).
The second step consists in combining all the $\widehat{\boldsymbol{\theta}}_{t,T}(m)$ for $m \in \{M/2^j, j = 0,\dots,k\}$ in the following way. Let $\boldsymbol{\alpha} = [\alpha_0 \ \dots \ \alpha_k]' \in \mathbb{R}^{k+1}$ be the solution of the equation $A\boldsymbol{\alpha} = \boldsymbol{e}_1$ where A is the $(k+1)\times(k+1)$ real matrix with entries $A_{i,j} = 2^{-(i-1)(j-1)}$ and $\boldsymbol{e}_1 = [1\ 0\ \dots\ 0]' \in \mathbb{R}^{k+1}$ has a 1 in the first component and zero everywhere else. Then, set $\widetilde{\boldsymbol{\theta}}_{t,T} = \sum_{j=0}^{k} \alpha_j \widehat{\boldsymbol{\theta}}_{t,T}(M/2^j)$.
Denote $\widehat{X}_{t,T} = \widehat{\boldsymbol{\theta}}'_{t,T} \boldsymbol{X}_{t-1:t-d,T}$, $\widehat{X}^*_{d,t,T} = (\boldsymbol{\theta}^*_{t,T})' \boldsymbol{X}_{t-1:t-d,T}$ and $\widehat{X}^*_{t,T} = \mathbb{E}[X_{t,T} | \sigma(X_{s,T}, s \le t-1)]$.
We obtain that, for T large enough and $q > 0$

$$\sup_{X\in\mathcal{M}_\lambda} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{\theta}}_{t,T}(M) - \boldsymbol{\theta}^*_{t,T}\right\|^q\right] \le C_1 \left(\frac{1}{M^{1/2}} + \left(\frac{M}{T}\right)^{\beta}\right)^q ,$$

and

$$\mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] \le \mathbb{E}\left[\left(\widehat{X}^*_{t,T} - X_{t,T}\right)^2\right] + \mathbb{E}\left[\left(\widehat{X}^*_{t,T} - \widehat{X}^*_{d,t,T}\right)^2\right] + C_2\left(\frac{1}{M^{1/2}} + \left(\frac{M}{T}\right)^{\beta}\right)^2 + C_3\left(\frac{1}{M^{1/2}} + \left(\frac{M}{T}\right)^{\beta}\right)\left(\mathbb{E}\left[\left(\widehat{X}^*_{t,T} - \widehat{X}^*_{d,t,T}\right)^2\right]\right)^{1/2} ,$$

where C_1, C_2 and C_3 depend only on $\boldsymbol{\lambda}$.

The result can be applied to locally stationary TVAR processes generated by $\boldsymbol{\theta} \in s_d(\delta) \cap \Lambda'_d(\beta, R)$ (with $\delta \in (0,1)$) and $\sigma \in \Lambda'(\beta, R) \cap [\rho\sigma_+, \sigma_+]^{(-\infty,1]}$. In this case we let $\boldsymbol{\lambda} = (\beta, R, \delta, \rho, \sigma_+)$ as in (2.3.17). Setting $M \propto T^{-2\beta/(2\beta+1)}$ we obtain the minimax rate for the regret

$$\sup_{X \in \mathcal{M}_\lambda} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] - \frac{1}{T} \sum_{t=1}^{T} \sigma^2 \left(\frac{t}{T} \right) \right\} \leq M_\lambda T^{-2\beta/(2\beta+1)} .$$

The NLMS based procedure to estimate $\boldsymbol{\theta}$ studied by Moulines et al. (2005) guaranties minimax convergence rates for $\beta \in (0, 2]$ on the TVAR model. We are not aware of any similar result for $\beta > 2$.

2.7 PERSPECTIVES

Aggregating an infinite number of predictors may pose the problem of computability as evoked in Chapter 3. When using Markov chain Monte Carlo methods, $M(T, \varepsilon)$, the number of iterations needed to achieve a numerical precision of the same order as the prediction risk may explode with the value of T (the horizon). In this context, scalable algorithm needs to be investigated.

Using a finite number of predictors, in Chapter 4 we study convex and best predicton regret upper bounds for sub-linear processes. In the particular case where the noise associated with the process and also the predictors have a finite pth-order moment, the optimality of these bounds remains an open problem.

A detailed analysis of aggregation strategies that do not rely on prior information on the process or its predictors should be conducted. It seems very likely that we could obtain such that crucial improvement without slowing down our oracle bounds rates.

Since TVAR process are sub-linear and locally stationary, the analysis of chapters 4 and 5 apply to them. When the order d is known, our contribution allows to propose an adaptive minimax predictor relying on the NLMS and the Yule-Walker algorithms and on the aggregation. In contrast, when d is unknown, it is not clear how to select the predictors to be aggregated.

When working with non-stationary processes, it is interesting to separate the roles of the number of observations or horizon (that we call T) and that of the sampling frequency ω (assumed to be T^{-1} through this thesis). A translation of our assumptions and results, from time series expressed as $(X_{t,T})_{1 \leq t \leq T}$ to time series expressed as $(X_{t,\omega})_{t \geq 1}$ may be a first step in this direction.

3 Time series prediction via aggregation: an oracle bound including numerical cost

Abstract

We address the problem of forecasting a time series meeting the Causal Bernoulli Shift model, using a parametric set of predictors. The aggregation technique provides a predictor with well established and quite satisfying theoretical properties expressed by an oracle inequality for the risk in prediction. The numerical computation of the aggregated predictor usually relies on a Markov chain Monte Carlo method whose convergence should be evaluated. In particular, it is crucial to bound the number of simulations needed to achieve a numerical precision of the same order as the risk in prediction. In this direction we present a fairly general result which can be seen as an oracle inequality including the numerical cost of the predictor computation. The numerical cost appears by letting the oracle inequality depend on the number of simulations required in the Monte Carlo approximation. Some numerical experiments are then carried out to support our findings.

3.1 Introduction

The objective of our work is to forecast a stationary time series $Y = (Y_t)_{t\in\mathbb{Z}}$ taking values in $\mathcal{X} \subseteq \mathbb{R}^r$ with $r \geq 1$. For this purpose we propose and study an aggregation scheme using exponential weights.

Consider a set of individual predictors giving their predictions at each moment t. An aggregation method consists of building a new prediction from this set, which is nearly as good as the best among the individual ones, provided a risk criterion (see Leung and Barron (2006)). This kind of result is established by oracle inequalities. The power and the beauty of the technique lie in its simplicity and versatility. The more basic and general context of application is individual sequences, where no assumption on the observations is made (see Cesa-Bianchi and Lugosi (2006) for a comprehensive overview). Nevertheless, results need to be adapted if we set a stochastic model on the observations.

The use of exponential weighting in aggregation and its links with the PAC-Bayesian approach has been investigated for example in Audibert (2004), Catoni (2004) and Dalalyan and Tsybakov (2008). Dependent processes have not received much attention from this viewpoint, except in Alquier and Li (2012) and Alquier and Wintenberger (2012). In the present paper we study the properties of the Gibbs predictor, applied to

Causal Bernoulli Shifts (CBS). CBS are an example of dependent processes (see Dedecker et al. (2007) and Dedecker and Prieur (2005)).
Our predictor is expressed as an integral since the set from which we do the aggregation is in general not finite. Large dimension is a trending setup and the computation of this integral is a major issue. We use classical Markov chain Monte Carlo (MCMC) methods to approximate it. Results from Łatuszyński Łatuszyński et al. (2013), Łatuszyński and Niemiro (2011) control the number of MCMC iterations to obtain precise bounds for the approximation of the integral. These bounds are in expectation and probability with respect to the distribution of the underlying Markov chain.
In this contribution we first slightly revisit certain lemmas presented in Alquier and Wintenberger (2012), Catoni (2004) and Rio (2000) to derive an oracle bound for the prediction risk of the Gibbs predictor. We stress that the inequality controls the convergence rate of the exact predictor. Our second goal is to investigate the impact of the approximation of the predictor on the convergence guarantees described for its exact version. Combining the PAC-Bayesian bounds with the MCMC control, we then provide an oracle inequality that applies to the MCMC approximation of the predictor, which is actually used in practice.
The paper is organised as follows: we introduce a motivating example and several definitions and assumptions in Section 3.2. In Section 3.3 we describe the methodology of aggregation and provide the oracle inequality for the exact Gibbs predictor. The stochastic approximation is studied in Section 3.4. We state a general proposition independent of the model for the Gibbs predictor. Next, we apply it to the more particular framework delineated in our paper. A concrete case study is analysed in Section 3.5, including some numerical work. A brief discussion follows in Section 3.6. The proofs of most of the results are deferred to Section 3.7.
Throughout the paper, for $\boldsymbol{a} \in \mathbb{R}^q$ with $q \in \mathbb{N}^*$, $\|\boldsymbol{a}\|$ denotes its Euclidean norm, $\|\boldsymbol{a}\| = (\sum_{i=1}^q a_i^2)^{1/2}$ and $\|\boldsymbol{a}\|_1$ its 1-norm $\|\boldsymbol{a}\|_1 = \sum_{i=1}^q |a_i|$. We denote, for $\boldsymbol{a} \in \mathbb{R}^q$ and $\Delta > 0$, $B(\boldsymbol{a}, \Delta) = \{\boldsymbol{a}_1 \in \mathbb{R}^q : \|\boldsymbol{a} - \boldsymbol{a}_1\| \leq \Delta\}$ and $B_1(\boldsymbol{a}, \Delta) = \{\boldsymbol{a}_1 \in \mathbb{R}^q : \|\boldsymbol{a} - \boldsymbol{a}_1\|_1 \leq \Delta\}$ the corresponding balls centered at $\boldsymbol{a}$ of radius $\Delta > 0$. In general bold characters represent column vectors and normal characters their components; for example $\boldsymbol{y} = (y_i)_{i\in\mathbb{Z}}$. The use of subscripts with ':' refers to certain vector components $\boldsymbol{y}_{1:k} = (y_i)_{1\leq i\leq k}$, or elements of a sequence $X_{1:k} = (X_t)_{1\leq t\leq k}$. For a random variable U distributed as ν and a measurable function h, $\nu[h(U)]$ or simply $\nu[h]$ stands for the expectation of $h(U)$: $\nu[h] = \int h(u)\nu(\mathrm{d}u)$.

3.2 Problem statement and main assumptions

Real stable autoregressive processes of a fixed order, referred to as AR(d) processes, are one of the simplest examples of CBS. They are defined as the stationary solution of

$$X_t = \sum_{j=1}^{d} \theta_j X_{t-j} + \sigma\xi_t \,, \tag{3.2.1}$$

where the $(\xi_t)_{t\in\mathbb{Z}}$ are i.i.d. real random variables with $\mathbb{E}[\xi_t] = 0$ and $\mathbb{E}[\xi_t^2] = 1$.

We dispose of several efficient estimates for the parameter $\boldsymbol{\theta} = [\theta_1 \ \dots \ \theta_d]'$ which can be calculated via simple algorithms as Levinson-Durbin or Burg algorithm for example. From them we derive also efficient predictors. However, as the model is simple to handle, we use it to progressively introduce our general setup.
Denote

$$A(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 & \theta_2 & \dots & \dots & \theta_d \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix},$$

$\boldsymbol{X}_{t-1} = [X_{t-1} \ \dots \ X_{t-d}]'$ and $\boldsymbol{e}_1 = [1\,0 \ \dots \ 0]'$ the first canonical vector of $\mathbb{R}^d$. M' represents the transpose of matrix M (including vectors). The recurrence (3.2.1) gives

$$X_t = \boldsymbol{\theta}' \boldsymbol{X}_{t-1} + \sigma\xi_t = \sigma \sum_{j=0}^{\infty} \boldsymbol{e}_1' A^j(\boldsymbol{\theta}) \boldsymbol{e}_1 \xi_{t-j} \,. \tag{3.2.2}$$

The eigenvalues of $A(\boldsymbol{\theta})$ are the inverses of the roots of the autoregressive polynomial $\boldsymbol{\theta}(z) = 1 - \sum_{k=1}^{d} \theta_k z^k$, then at most δ for some $\delta \in (0,1)$ due to the stability of X (see Brockwell and Davis (2006)). In other words $\boldsymbol{\theta} \in s_d(\delta) = \{\boldsymbol{\theta} \ : \ \boldsymbol{\theta}(z) \neq 0 \text{ for } |z| < \delta^{-1}\} \subseteq s_d(1)$. In this context (or even in a more general one, see Künsch (1995)) for all $\delta_1 \in (\delta, 1)$ there is a constant $\bar{K}$ depending only on $\boldsymbol{\theta}$ and δ_1 such that for all $j \geq 0$

$$\left|\boldsymbol{e}_1' A^j(\boldsymbol{\theta}) \boldsymbol{e}_1\right| \leq \bar{K}\delta_1^j \,, \tag{3.2.3}$$

and then, the variance of X_t, denoted γ_0, satisfies $\gamma_0 = \sigma^2 \sum_{j=0}^{\infty} |\boldsymbol{e}_1' A^j(\boldsymbol{\theta}) \boldsymbol{e}_1|^2 \leq \bar{K}^2\sigma^2/(1-\delta_1^2)$.
The following definition allows to introduce the process which interests us.

Definition 5. *Let $\mathcal{X}' \subseteq \mathbb{R}^{r'}$ for some $r' \geq 1$ and let $A = (A_j)_{j\geq 0}$ be a sequence of non-negative numbers. A function $H : (\mathcal{X}')^{\mathbb{N}} \to \mathcal{X}$ is said to be A-Lipschitz if*

$$\|H(\boldsymbol{u}) - H(\boldsymbol{v})\| \ \leq \ \sum_{j=0}^{\infty} A_j \|u_j - v_j\| \,,$$

for any $\boldsymbol{u} = (u_j)_{j\in\mathbb{N}}, \boldsymbol{v} = (v_j)_{j\in\mathbb{N}} \in (\mathcal{X}')^{\mathbb{N}}$.

Provided $A = (A_j)_{j\geq 0}$ with $A_j \geq 0$ for all $j \geq 0$, the i.i.d. sequence of $\mathcal{X}'$-valued random variables $(\xi_t)_{t\in\mathbb{Z}}$ and $H : (\mathcal{X}')^{\mathbb{N}} \to \mathcal{X}$, we consider that a time series $X = (X_t)_{t\in\mathbb{Z}}$ admitting the following property is a Causal Bernoulli Shift (CBS) with Lipschitz coefficients A and innovations $(\xi_t)_{t\in\mathbb{Z}}$.

(**M**) The process $X = (X_t)_{t\in\mathbb{Z}}$ meets the representation

$$X_t \ = \ H(\xi_t, \xi_{t-1}, \xi_{t-2}, \dots) \,, \forall t \in \mathbb{Z} \,,$$

where H is an A-Lipschitz function with the sequence A satisfying

$$\tilde{A}_* = \sum_{j=0}^{\infty} jA_j < \infty \,. \tag{3.2.4}$$

We additionally define

$$A_* = \sum_{j=0}^{\infty} A_j \,. \tag{3.2.5}$$

CBS regroup several types of nonmixing stationary Markov chains, real-valued functional autoregressive models and Volterra processes, among other interesting models (see Coulon-Prieur and Doukhan (2000)). Thanks to the representation (3.2.2) and the inequality (3.2.3) we assert that AR(d) processes are CBS with $A_j = \sigma \bar{K} \delta_1^j$ for $j \geq 0$.
We let ξ denote a random variable distributed as the ξ_ts. Results from Alquier and Li (2012) and Alquier and Wintenberger (2012) need a control on the exponential moment of ξ in $\zeta = A_*$, which is provided via the following hypothesis.

(**I**) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ satisfy $\phi(\zeta) = \mathbb{E}\left[\mathrm{e}^{\zeta\|\xi\|}\right] < \infty$.

Bounded or Gaussian innovations trivially satisfy this hypothesis for any $\zeta \in \mathbb{R}$.
Let υ denote the probability distribution of the time series Y that we aim to forecast. Observe that for a CBS, υ depends only on H and the distribution of ξ. For any $f : \mathcal{X}^{\mathbb{N}^*} \to \mathcal{X}$ measurable and $t \in \mathbb{Z}$ we consider $\widehat{Y}_t = f\left((Y_{t-i})_{i\geq 1}\right)$, a possible predictor of Y_t from its past. For a given loss function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the risk is evaluated by the expectation of $\ell(\widehat{Y}_t, Y_t)$

$$R(f) = \mathbb{E}\left[\ell\left(\widehat{Y}_t, Y_t\right)\right] = \upsilon\left[\ell\left(\widehat{Y}_t, Y_t\right)\right] = \int_{\mathcal{X}^{\mathbb{Z}}} \ell\left(f\left((y_{t-i})_{i\geq 1}\right), y_t\right) \upsilon(\mathrm{d}\boldsymbol{y}) \,.$$

We assume in the following that the loss function ℓ fulfills the condition:

(**L**) For all $\boldsymbol{y}, \boldsymbol{z} \in \mathcal{X}$, $\ell(\boldsymbol{y}, \boldsymbol{z}) = g(\boldsymbol{y} - \boldsymbol{z})$, for some convex function g which is non-negative, $g(0) = 0$ and K- Lipschitz: $|g(\boldsymbol{y}) - g(\boldsymbol{z})| \leq K\|\boldsymbol{y} - \boldsymbol{z}\|$.

If $\mathcal{X}$ is a subset of $\mathbb{R}$, $\ell(y, z) = |y - z|$ satisfies (**L**) with $K = 1$. If $\mathcal{X} \subset \mathbb{R}$ is bounded, this is, there exists $B > 0$ such that $\|x\| \leq B$ for all $x \in \mathcal{X}$, the quadratic loss meets Assumption (**L**) with $K = 2B$.
From estimators of dimension d for $\boldsymbol{\theta}$ we can build the corresponding linear predictors $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}' \boldsymbol{y}_{1:d}$. Speaking more broadly, consider a set Θ and associate with it a set of predictors $\{f_\theta, \boldsymbol{\theta} \in \Theta\}$. For each $\boldsymbol{\theta} \in \Theta$ there is a unique $d = d(\boldsymbol{\theta}) \in \mathbb{N}^*$ such that $f_{\boldsymbol{\theta}} : \mathcal{X}^d \to \mathcal{X}$ is a measurable function from which we define

$$\widehat{Y}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(Y_{t-1}, \ldots, Y_{t-d}) \,,$$

as a predictor of Y_t given its past. We can extend all functions $f_{\boldsymbol{\theta}}$ in a trivial way (using dummy variables) to start from $\mathcal{X}^{\mathbb{N}^*}$. A natural way to evaluate the predictor associated

with $\boldsymbol{\theta}$ is to compute the risk $R(\boldsymbol{\theta}) = R(f_{\boldsymbol{\theta}})$. We use the same letter R by an abuse of notation.
Given $\widehat{Y}_t = f((Y_{t-i})_{i\geq 1})$ and a σ-field $\mathcal{H}$, we define the prediction risk by

$$R(f|\mathcal{H}) = \mathbb{E}\left[\ell\left(\widehat{Y}_t, Y_t\right)\middle|\mathcal{H}\right] . \tag{3.2.6}$$

In the case where $\mathcal{H} = \sigma(U)$ for some random variable U, we write $R(f|U)$.
We observe $X_{1:T}$ from $X = (X_t)_{t\in\mathbb{Z}}$, an independent copy of Y. A crucial goal of this work is to build a predictor function $\hat{f}_T$ for Y, inferred from the sample $X_{1:T}$ and Θ such that the prediction risk $R(\hat{f}_T|X)$ is close to $\inf_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta})$ with υ- probability close to 1.
The set Θ also depends on T, we write $\Theta \equiv \Theta_T$. Let us define

$$d_T = \sup_{\boldsymbol{\theta}\in\Theta_T} d(\boldsymbol{\theta}) . \tag{3.2.7}$$

The main assumptions on the set of predictors are the following ones.

(**P**-1) The set $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta}\in\Theta_T\}$ is such that for any $\boldsymbol{\theta}\in\Theta_T$ there are $b_1(\boldsymbol{\theta}), \ldots, b_{d(\boldsymbol{\theta})}(\boldsymbol{\theta}) \in \mathbb{R}_+$ satisfying for all $\boldsymbol{y} = (y_i)_{i\in\mathbb{N}^*}, \boldsymbol{z} = (z_i)_{i\in\mathbb{N}^*} \in \mathcal{X}^{\mathbb{N}^*}$,

$$\|f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\boldsymbol{z})\| \leq \sum_{j=1}^{d(\boldsymbol{\theta})} b_j(\boldsymbol{\theta}) \left\|y_j - z_j\right\| .$$

We assume moreover that $L_T = \sup_{\boldsymbol{\theta}\in\Theta_T} \sum_{j=1}^{d(\boldsymbol{\theta})} b_j(\boldsymbol{\theta}) < \infty$.

(**P**-2) The inequality $L_T + 1 \leq \log T$ holds for all $T \geq 4$.

In the case where $\mathcal{X}\subseteq\mathbb{R}$ and $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta}\in\Theta_T\}$ is such that $\boldsymbol{\theta}\in\mathbb{R}^{d(\boldsymbol{\theta})}$ and $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}'\boldsymbol{y}_{1:d(\boldsymbol{\theta})}$ for all $\boldsymbol{y}\in\mathbb{R}^{\mathbb{N}}$, we have

$$|f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\boldsymbol{\theta}}(\boldsymbol{z})| \leq \sum_{j=1}^{d(\boldsymbol{\theta})} |\theta_j| \left|y_j - z_j\right| .$$

The last conditions are satisfied by the linear predictors when Θ_T is a subset of the ℓ_1-ball of radius $\log T - 1$ in $\mathbb{R}^{d_T}$.

3.3 PREDICTION VIA AGGREGATION

The predictor that we propose is defined as an average of predictors $f_{\boldsymbol{\theta}}$ based on the empirical version of the risk,

$$r_T(\boldsymbol{\theta}|X) = \frac{1}{T - d(\boldsymbol{\theta})} \sum_{t=d(\boldsymbol{\theta})+1}^{T} \ell\left(\widehat{X}_t^{\boldsymbol{\theta}}, X_t\right) .$$

where $\widehat{X}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}((X_{t-i})_{i\geq 1})$. The function $r_T(\boldsymbol{\theta}|X)$ relies on $X_{1:T}$ and can be computed at stage T; this is in fact a statistic.
We consider a prior probability measure π_T on Θ_T. The prior serves to control the complexity of predictors associated with Θ_T. Using π_T we can construct one predictor in particular, as detailed in the following.

3.3.1 Gibbs predictor

For a measure ν and a measurable function h (called energy function) such that $\nu[\exp(h)] = \int \exp(h)\ \mathrm{d}\nu < \infty$, we denote by $\nu\{h\}$ the measure defined as

$$\nu\{h\}(\mathrm{d}\boldsymbol{\theta}) = \frac{\exp(h(\boldsymbol{\theta}))}{\nu[\exp(h)]}\nu(\mathrm{d}\boldsymbol{\theta}) \ .$$

It is known as the Gibbs measure.

Definition 6 (Gibbs predictor)**.** *Given $\eta > 0$, called the temperature or the learning rate parameter, we define the Gibbs predictor as the expectation of $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is drawn under $\pi_T\{-\eta r_T(\cdot|X)\}$, that is*

$$\hat{f}_{\eta,T}(\boldsymbol{y}|X) = \pi_T\{-\eta r_T(\cdot|X)\}[f_\cdot(\boldsymbol{y})] = \int_{\Theta_T} f_{\boldsymbol{\theta}}(\boldsymbol{y})\frac{\exp(-\eta r_T(\boldsymbol{\theta}|X))}{\pi_T[\exp(-\eta r_T(\cdot|X))]}\pi_T(\mathrm{d}\boldsymbol{\theta}) \ . \quad (3.3.1)$$

3.3.2 PAC-Bayesian inequality

At this point more care must be taken to describe Θ_T. Here and in the following we suppose that

$$\Theta_T \subseteq \mathbb{R}^{n_T} \ \text{ for some } n_T \in \mathbb{N}^* \ . \quad (3.3.2)$$

Suppose moreover that Θ_T is equipped with the Borel σ-algebra $\mathcal{B}(\Theta_T)$.
A Lipschitz type hypothesis on $\boldsymbol{\theta}$ guarantees the robustness of the set $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ with respect to the risk R.

(**P**-3) There is $\mathcal{D} < \infty$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_T$,

$$\upsilon\left[\left\|f_{\boldsymbol{\theta}_1}((X_{t-i})_{i\geq 1}) - f_{\boldsymbol{\theta}_2}((X_{t-i})_{i\geq 1})\right\|\right] \leq \mathcal{D} d_T^{1/2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \ ,$$

where d_T is defined in (3.2.7).

Linear predictors satisfy this last condition with $\mathcal{D} = \upsilon[|X_1|]$.
Suppose that the $\boldsymbol{\theta}$ reaching the $\inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta})$ has some zero components, i.e. $\mathrm{supp}(\boldsymbol{\theta}) < n_T$. Any prior with a lower bounded density (with respect to the Lebesgue measure) allocates zero mass on lower dimensional subsets of Θ_T. Furthermore, if the density is upper bounded we have $\pi_T[B(\boldsymbol{\theta}, \Delta) \cap \Theta_T] = O(\Delta^{n_T})$ for Δ small enough. As we will notice in the proof of Theorem 3.3.1, a bound like the previous one would impose a tighter constraint to n_T. Instead we set the following condition.

(**P**-4) There is a sequence $(\boldsymbol{\theta}_T)_{T\geq 4}$ and constants $C_1 > 0$, $C_2, C_3 \in (0, 1]$ and $\gamma \geq 1$ such that $\boldsymbol{\theta}_T \in \Theta_T$,

$$R(\boldsymbol{\theta}_T) \leq \inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta}) + C_1\frac{\log^3(T)}{T^{1/2}} \ ,$$

$$\text{and} \quad \pi_T[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T] \geq C_2\Delta^{n_T^{1/\gamma}}, \forall 0 \leq \Delta \leq \Delta_T = \frac{C_3}{T} \ .$$

A concrete example is provided in Section 3.5.
We can now present the main result of this section, our PAC-Bayesian inequality concerning the predictor $\hat{f}_{\eta_T,T}(\cdot|X)$ built following (3.3.1) with the learning rate $\eta = \eta_T = T^{1/2}/(4\log T)$, provided an arbitrary probability measure π_T on Θ_T.

Theorem 3.3.1. *Let ℓ be a loss function such that Assumption (L) holds. Consider a process $X = (X_t)_{t\in\mathbb{Z}}$ satisfying Assumption (M) and let υ denote its probability distribution. Assume that the innovations fulfill Assumption (I) with $\zeta = A_*$; A_* is defined in (3.2.5). For each $T \geq 4$ let $\{f_\theta, \theta \in \Theta_T\}$ be a set of predictors meeting Assumptions (P-1), (P-2) and (P-3) such that d_T, defined in (3.2.7), is at most $T/2$. Suppose that the set Θ_T is as in (3.3.2) with $n_T \leq \log^\gamma T$ for some $\gamma \geq 1$ and we let π_T be a probability measure on it such that Assumption (P-4) holds for the same γ. Then for any $\varepsilon > 0$, with υ-probability at least $1-\varepsilon$,*

$$R\left(\hat{f}_{\eta_T,T}(\cdot|X)\middle|X\right) \leq \inf_{\theta\in\Theta_T} R(f_\theta) + \mathcal{E}\frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}}\log\left(\frac{1}{\varepsilon}\right),$$

where

$$\mathcal{E} = C_1 + 8 + \frac{2}{\log 2} - \frac{2\log C_2}{\log^2 2} - \frac{4\log C_3}{\log 2} + \frac{8K^2\left(A_* + \tilde{A}_*\right)^2}{\tilde{A}_*^2} + \frac{K\mathcal{D}C_3}{8\log^3 2} + \frac{4K\phi(A_*)}{\log 2} + \frac{2K^2\phi(A_*)}{\log^2 2}, \tag{3.3.3}$$

with $\tilde{A}_*$ defined in (3.2.4), K, ϕ and $\mathcal{D}$ in Assumptions (L), (I) and (P-3), respectively, and C_1, C_2 and C_3 in Assumption (P-4).

The proof is postponed to Section 3.7.1.
Here however we insist on the fact that this inequality applies to an exact aggregated predictor $\hat{f}_{\eta_T,T}(\cdot|X)$. We need to investigate how these predictors are computed and how practical numerical approximations behave compared to the properties of the exact version.

3.4 Stochastic approximation

Once we have the observations $X_{1:T}$, we use the Metropolis - Hastings algorithm to compute $\hat{f}_{\eta,T}(\cdot|X) = \int f_\theta(\cdot|X)\,\pi_T\{-\eta r_T(\theta|X)\}(\mathrm{d}\theta)$. The Gibbs measure $\pi_T\{-\eta r_T(\cdot|X)\}$ is a distribution on Θ_T whose density $\pi_{\eta,T}(\cdot|X)$ with respect to π_T is proportional to $\exp(-\eta r_T(\cdot|X))$.

3.4.1 Metropolis - Hastings algorithm

Given $X \in \mathcal{X}^{\mathbb{Z}}$, the Metropolis-Hastings algorithm generates a Markov chain $\Phi_{\eta,T}(X) = (\theta_{\eta,T,n}(X))_{n\geq 0}$ with kernel $P_{\eta,T}$ (only depending on $X_{1:T}$) having the target distribution

$\pi_T\{-\eta r_T(\cdot|X)\}$ as the unique invariant measure, based on the transitions of another Markov chain which serves as a proposal (see Roberts and Rosenthal (2004)). We consider a proposal transition of the form $Q_{\eta,T}(\theta_1, d\theta) = q_{\eta,T}(\theta_1, \theta)\pi_T(d\theta)$ where the conditional density kernel $q_{\eta,T}$ (possibly also depending on $X_{1:T}$) on $\Theta_T \times \Theta_T$ is such that

$$\beta_{\eta,T}(X) = \inf_{(\theta_1,\theta_2)\in\Theta_T\times\Theta_T} \frac{q_{\eta,T}(\theta_1,\theta_2)}{\pi_{\eta,T}(\theta_2|X)} \in (0,1) \ . \tag{3.4.1}$$

This is the case of the independent Hastings algorithm, where the proposal is i.i.d. with density $q_{\eta,T}$. The condition gets into

$$\beta_{\eta,T}(X) = \inf_{\theta\in\Theta_T} \frac{q_{\eta,T}(\theta)}{\pi_{\eta,T}(\theta|X)} \in (0,1) \ . \tag{3.4.2}$$

In Section 3.5 we provide an example.
The relation (3.4.1) implies that the algorithm is uniformly ergodic, i.e. we have a control in total variation norm ($\|\cdot\|_{TV}$). Thus, the following condition holds (see Mengersen and Tweedie (1996)).

(**A**) Given $\eta, T > 0$, there is $\beta_{\eta,T} : \mathcal{X}^{\mathbb{Z}} \to (0,1)$ such for any $\theta_0 \in \Theta_T$, $x \in \mathcal{X}^{\mathbb{Z}}$ and $n \in \mathbb{N}$, the chain $\Phi_{\eta,T}(x)$ with transition law $P_{\eta,T}$ and invariant distribution $\pi_T\{-\eta r_T(\cdot|x)\}$ satisfies

$$\left\|P^n_{\eta,T}(\theta_0,\cdot) - \pi_T\{-\eta r_T(\cdot|x)\}\right\|_{TV} \ \leq \ 2\left(1-\beta_{\eta,T}(x)\right)^n \ .$$

3.4.2 Theoretical bounds for the computation

In (Łatuszyński and Niemiro, 2011, Theorem 3.1) we find a bound on the mean square error of approximating one integral by the empirical estimate obtained from the successive samples of certain ergodic Markov chains, including those generated by the MCMC method that we use.
A MCMC method adds a second source of randomness to the forecasting process and our aim is to measure it. Let $\theta_0 \in \cap_{T\geq 1}\Theta_T$, we set $\theta_{\eta,T,0}(x) = \theta_0$ for all $T, \eta > 0$, $x \in \mathcal{X}^{\mathbb{Z}}$. We denote by $\mu_{\eta,T}(\cdot|X)$ the probability distribution of the Markov chain $\Phi_{\eta,T}(X)$ with initial point θ_0 and kernel $P_{\eta,T}$.
Let $\nu_{\eta,T}$ denote the probability distribution of $(X, \Phi_{\eta,T}(X))$; it is defined by setting for all sets $A \in (\mathcal{B}(\mathcal{X}))^{\otimes\mathbb{Z}}$ and $B \in (\mathcal{B}(\Theta_T))^{\otimes\mathbb{N}}$

$$\nu_{\eta,T}(A\times B) = \int \mathbb{1}_A(x)\,\mathbb{1}_B(\phi)\,\mu_{\eta,T}(d\phi|x)\,\upsilon(dx) \tag{3.4.3}$$

Given $\Phi_{\eta,T} = (\theta_{\eta,T,n})_{n\geq 0}$, we then define for $n \in \mathbb{N}^*$

$$\bar{f}_{\eta,T,n} = \frac{1}{n}\sum_{i=0}^{n-1} f_{\theta_{\eta,T,i}} \ . \tag{3.4.4}$$

Since our chain depends on X, we make it explicit by using the notation $\bar{f}_{\eta,T,n}(\cdot\,|X)$. The cited (Łatuszyński and Niemiro, 2011, Theorem 3.1) leads to a proposition that applies to the numerical approximation of the Gibbs predictor (the proof is in Section 3.7.2). We stress that this is independent of the model (CBS or any), of the set of predictors and of the theoretical guarantees of Theorem 3.3.1.

Proposition 1. *Let ℓ be a loss function meeting Assumption (L). Consider any process $X = (X_t)_{t\in\mathbb{Z}}$ with an arbitrary probability distribution υ. Given $T \geq 2$, $\eta > 0$, a set of predictors $\{f_\theta, \theta \in \Theta_T\}$ and $\pi_T \in \mathcal{M}_+^1(\Theta_T)$, let $\hat{f}_{\eta,T}(\cdot\,|X)$ be defined by (3.3.1) and let $\bar{f}_{\eta,T,n}(\cdot\,|X)$ be defined by (3.4.4). Suppose that $\Phi_{\eta,T}$ meets Assumption (A) for η and T with a function $\beta_{\eta,T} : \mathcal{X}^{\mathbb{Z}} \to (0,1)$. Let $\nu_{\eta,T}$ denote the probability distribution of $(X, \Phi_{\eta,T}(X))$ as defined in (3.4.5). Then, for all $n \geq 1$ and $D > 0$, with $\nu_{\eta,T}$- probability at least $\max\{0, 1 - A_{\eta,T}/(Dn^{1/2})\}$ we have $|R(\bar{f}_{\eta,T,n}(\cdot\,|X)\,|X, \Phi_{\eta,T}) - R(\hat{f}_{\eta,T}(\cdot\,|X)\,|X, \Phi_{\eta,T})| \leq D$, where*

$$A_{\eta,T} = 3K \int_{\mathcal{X}^{\mathbb{Z}}} \frac{1}{\beta_{\eta,T}(\boldsymbol{x})} \int_{\mathcal{X}^{\mathbb{Z}}} \sup_{\theta\in\Theta_T} \left| f_\theta(\boldsymbol{y}) - \hat{f}_{\eta,T}(\boldsymbol{y}\,|\boldsymbol{x}) \right| \upsilon(\mathrm{d}\boldsymbol{y})\, \upsilon(\mathrm{d}\boldsymbol{x}) \,. \tag{3.4.5}$$

We denote by $\nu_T = \nu_{\eta_T,T}$ the probability distribution of $(X, \Phi_{\eta,T}(X))$ setting $\eta = \eta_T = T^{1/2}/(4\log T)$. As Theorem 3.3.1 does not involve any simulation, it also holds in ν_T- probability. From this and Proposition 1 a union bound gives us the following.

Theorem 3.4.1. *Under the hypothesis of Theorem 3.3.1, consider moreover that Assumption (A) is fulfilled by $\Phi_{\eta,T}$ for all $\eta = \eta_T$ and T with $T \geq 4$. Thus, for all $\varepsilon > 0$ and $n \geq M(T,\varepsilon)$, with ν_T- probability at least $1 - \varepsilon$ we have*

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot\,|X)\middle|\, X, \Phi_{\eta_T,T}\right) \leq \inf_{\theta\in\Theta_T} R(f_\theta) + \left(\mathcal{E} + \frac{2}{\log 2} + 2\right)\frac{\log^3 T}{T^{1/2}} + \frac{8\log T}{T^{1/2}} \log\left(\frac{1}{\varepsilon}\right),$$

where $\mathcal{E}$ is defined in (3.3.3) and $M(T,\varepsilon) = A_{\eta_T,T}^2 T/(\varepsilon^2 \log^6 T)$ with $A_{\eta,T}$ as in (3.4.5).

Theorem 3.4.1 recalls the work of Kalai and Vempala (2002). They propose an efficient implementation of Cover's universal algorithm for portfolios (see Cover (1991)). The mentioned algorithm relies on the evaluation of an integral. The method of Kalai and Vempala (2002) introduces a randomized approximation of it and exhibits guaranties with high probability.

3.5 APPLICATIONS TO THE AUTOREGRESSIVE PROCESS

We carefully recapitulate all the assumptions of Theorem 3.4.1 in the context of an autoregressive process. After that, we illustrate numerically the behaviour of the proposed method.

3.5.1 Theoretical considerations

Consider a real valued stable autoregressive process of finite order d as defined by (3.2.1) with parameter $\boldsymbol{\theta}$ lying in the interior of $s_d(\delta)$ and unit normally distributed innovations (Assumptions **(M)** and **(I)** hold). With the loss function $\ell(y,z) = |y-z|$ Assumption **(L)** holds as well. The linear predictors is the set that we test; they meet Assumption **(P-3)**. Without loss of generality assume that $d_T = n_T$. In the described framework we have $\hat{f}_{\eta,T}(\cdot|X) = f_{\hat{\boldsymbol{\theta}}_{\eta,T}(X)}$, where

$$\hat{\boldsymbol{\theta}}_{\eta,T}(X) = \int_{\Theta_T} \boldsymbol{\theta} \frac{\exp\left(-\eta r_T(\boldsymbol{\theta}|X)\right)}{\pi_T\left[\exp\left(-\eta r_T(\boldsymbol{\theta}|X)\right)\right]} \pi_T(\mathrm{d}\boldsymbol{\theta}) \ .$$

This $\hat{\boldsymbol{\theta}}_{\eta,T}(X) \in \mathbb{R}^{d_T}$ is known as the Gibbs estimator.
Remark that, by (3.2.2) and the normality of the innovations, the risk of any $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{d_T}$ is computed as the absolute moment of a centered Gaussian, namely

$$R(f_{\hat{\theta}}) = R\left(\hat{\boldsymbol{\theta}}\right) = \frac{\left(2\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)' \Gamma_T \left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) + 2\sigma^2\right)^{1/2}}{\pi^{1/2}} \ , \tag{3.5.1}$$

where $\Gamma_T = (\gamma_{i,j})_{0\le i,j\le d_T-1}$ is the covariance matrix of the process. In (3.5.1) the vector $\boldsymbol{\theta}$ originally in $\mathbb{R}^d$ is completed by $d_T - d$ zeros.
In this context $\arg\inf_{\boldsymbol{\theta}\in\mathbb{R}^{\mathbb{N}^*}} R(\boldsymbol{\theta}) \in s_d(1)$ gives the true parameter $\boldsymbol{\theta}$ generating the process. Let us verify Assumption **(P-4)** by setting conveniently Θ_T and π_T. Let $\Delta_{d*} > 0$ be such that $B(\boldsymbol{\theta}, \Delta_{d*}) \subseteq s_d(1)$.
We express $\Theta_T = \bigcup_{k=1}^{d_T} \Theta_{k,T}$ where $\boldsymbol{\theta} \in \Theta_{k,T}$ if and only if $d(\boldsymbol{\theta}) = k$. It is interesting to set $\Theta_{k,T}$ as the part of the stability domain of an AR(k) process satisfying Assumptions **(P-1)** and **(P-2)**. Consider $\Theta_{1,T} = s_1(1) \times \{0\}^{d_T-1} \cap B_1(\mathbf{0}, \log T - 1)$ and $\Theta_{k,T} = s_k(1) \times \{0\}^{d_T-k} \cap B_1(\mathbf{0}, \log T - 1) \setminus (\cup_{k'=1}^{k-1} \Theta_{k',T})$ for $k \ge 2$. Assume moreover that $d_T = \lfloor \log^\gamma T \rfloor$.
We write $\pi_T = \sum_{k=1}^{d_T} c_{k,T}\pi_{k,T}$ where for all k, $c_{k,T}\pi_{k,T}$ is the restriction of π_T to $\Theta_{k,T}$ with $c_{k,T}$ a real non negative number and $\pi_{k,T}$ a probability measure on $\Theta_{k,T}$. In this setup $c_{k,T} = \pi_T[\Theta_{k,T}]$ and $\pi_{k,T}[A \cap \Theta_{k,T}] = \pi_T[A \cap \Theta_{k,T}]/c_{k,T}$ if $c_{k,T} > 0$ and $\pi_{k,T}[A \cap \Theta_{k,T}] = 0$ otherwise. The vector $[c_{1,T} \dots c_{d_T,T}]$ could be interpreted as a prior on the model order. Set $c_{k,T} = c_k / (\sum_{i=1}^{d_T} c_i)$ where $c_k > 0$ is the k-th term of a convergent series ($\sum_{k=1}^{\infty} c_k = c^* < \infty$). The distribution $\pi_{k,T}$ is inferred from some transformations explained below. Observe first that if $a \le b$ we have $s_k(a) \subseteq s_k(b)$. If $\boldsymbol{\theta} \in s_k(1)$ then $[\lambda\theta_1 \dots \lambda^k\theta_k]' \in s_k(1)$ for any $\lambda \in (-1,1)$. Let us set

$$\lambda_T(\boldsymbol{\theta}) = \min\left\{1, \frac{\log T - 1}{\|\boldsymbol{\theta}\|_1}\right\} \ .$$

We define $F_{k,T}(\boldsymbol{\theta}) = [\lambda_T(\boldsymbol{\theta})\theta_1 \dots \lambda_T^k(\boldsymbol{\theta})\theta_k\, 0 \dots 0]' \in \mathbb{R}^{d_T}$. Remark that for any $\boldsymbol{\theta} \in s_k(1)$, $\|F_{k,T}(\boldsymbol{\theta})\|_1 \le \lambda_T(\boldsymbol{\theta})\|\boldsymbol{\theta}\|_1 \le \log T - 1$. This gives us an idea to generate vectors in $\Theta_{k,T}$. Our distribution $\pi_{k,T}$ is deduced from:

Algorithm 3: $\pi_{k,T}$ generation

input an effective dimension k, the number of observations T and $F_{k,T}$;
generate a random $\boldsymbol{\theta}$ uniformly on $s_k(1)$;
return $F_{k,T}(\boldsymbol{\theta})$

The distribution $\pi_{k,T}$ is lower bounded by the uniform distribution on $s_k(1)$.
Provided any $\gamma \geq 1$, let $T_* = \min\{T : d_T \geq d^\gamma, \log T \geq d^{1/2}2^d\}$. Since $s_k(1) \subseteq B(\mathbf{0}, 2^k - 1)$ (see (Moulines et al., 2005, Lemma 1)) and $k^{1/2}\|\boldsymbol{\theta}\| \geq \|\boldsymbol{\theta}\|_1$ for any $\boldsymbol{\theta} \in \mathbb{R}^k$, the constraint $\|\boldsymbol{\theta}\|_1 \leq \log T - 1$ becomes redundant in $\Theta_{k,T}$ for $1 \leq k \leq d$ and $T \geq T_*$, i.e. $\Theta_{1,T} = s_1(1) \times \{0\}^{d_T-1}$ and $\Theta_{k,T} = s_k(1) \times \{0\}^{d_T-k} \backslash \Theta_{k-1,T}$ for $2 \leq k \leq d$. We define the sequence of Assumption (P-4) as $\boldsymbol{\theta}_T = \mathbf{0}$ for $T < T_*$ and $\boldsymbol{\theta}_T = \arg\inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta})$ for $T \geq T_*$. Remark that the first d components of $\boldsymbol{\theta}_T$ are constant for $T \geq T_*$ (they correspond to the $\boldsymbol{\theta} \in \mathbb{R}^d$ generating the AR(d) process), and the last $d_T - d$ are zero. Let $\Delta_{1*} = 2\log 2 - 1$. Then, we have for $T < T_*$ and all $\Delta \in [0, \Delta_{1*}]$

$$\pi_T\left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T\right] \geq c_{1,T}\pi_{1,T}\left[B(\mathbf{0}, \Delta) \cap s_1(1) \times \{0\}^{d_T-1}\right] \geq \frac{c_1}{c^*}\Delta\,.$$

Furthermore, for $T \geq T_*$ and $\Delta \in [0, \Delta_{d*}]$

$$\pi_T\left[B(\boldsymbol{\theta}_T, \Delta) \cap \Theta_T\right] \geq c_{d,T}\pi_{d,T}\left[B(\boldsymbol{\theta}_T, \Delta) \cap s_d(1) \times \{0\}^{d_T-d}\right] \geq \frac{c_d}{2^{d^2}c^*}\Delta^d\,.$$

Assumption (P-4) is then fulfilled for any $\gamma \geq 1$ with

$$\begin{aligned} C_1 &= \max\left\{0, (R(0) - \inf_{\boldsymbol{\theta}\in\Theta_T} R(\boldsymbol{\theta}))T^{1/2}\log^{-3} T, 4 \leq T < T_*\right\} \\ C_2 &= \min\left\{1, \frac{c_1}{c^*}, \frac{c_d}{2^{d^2}c^*}\right\} \\ C_3 &= \min\{1, 4\Delta_{1*}, T_*\Delta_{d*}\}\,. \end{aligned}$$

Let $q_{\eta,T}$ be the constant function 1, this means that the proposal has the same distribution π_T. Let us bound the ratio (3.4.2).

$$\begin{aligned} \beta_{\eta,T}(X) = \inf_{\boldsymbol{\theta}\in\Theta_T}\frac{q_{\eta,T}(\boldsymbol{\theta})}{\pi_{\eta,T}(\boldsymbol{\theta}|X)} &= \inf_{\boldsymbol{\theta}\in\Theta_T}\frac{\sum_{k=1}^{d_T} c_{k,T}\int_{\Theta_{k,T}}\exp(-\eta r_T(z|X))\,\pi_{k,T}(\mathrm{d}z)}{\exp(-\eta r_T(\boldsymbol{\theta}|X))} \\ &\geq \sum_{k=1}^{d_T} c_{k,T}\int_{\Theta_{k,T}}\exp(-\eta r_T(z|X))\,\pi_{k,T}(\mathrm{d}z) > 0\,. \end{aligned} \quad (3.5.2)$$

Now note that

$$\left|x_t - f_{\boldsymbol{\theta}}((x_{t-i})_{i\geq 1})\right| \leq |x_t| + \sum_{j=1}^{d(\boldsymbol{\theta})}\left|\theta_j\right|\left|x_{t-j}\right| \leq \log T \max_{j=0,\ldots,d(\boldsymbol{\theta})}\left|x_{t-j}\right|\,. \quad (3.5.3)$$

Plugging the bound (3.5.3) on (3.5.2) with $\eta = \eta_T$

$$\beta_{\eta_T,T}(\boldsymbol{x}) \geq \sum_{k=1}^{d_T} c_k \int_{\Theta_k} \exp\left(-\eta_T r_T(z|\boldsymbol{x})\right) \pi_k(\mathrm{d}z) \geq \exp\left(-\frac{T^{1/2}}{4} \max_{j=0,\ldots,d_T} |x_{t-j}|\right),$$

we deduce that

$$\frac{1}{\beta_{\eta_T,T}(\boldsymbol{x})} \leq \sum_{k=0}^{d_T} \exp\left(\frac{T^{1/2}|x_{t-j}|}{4}\right). \tag{3.5.4}$$

Taking (3.5.4) into account, setting $\gamma = 1$ (thus $d_T = \lfloor \log T \rfloor$), using Assumption (**P**-3), that $K = 1$ and applying the Cauchy-Schwarz inequality we get

$$\begin{aligned} A_{\eta_T,T} &= 3K \int_{\mathcal{X}^{\mathbb{Z}}} \frac{1}{\beta_{\eta_T,T}(\boldsymbol{x})} \int_{\mathcal{X}^{\mathbb{Z}}} \sup_{\boldsymbol{\theta}\in\Theta_T} \left| f_{\boldsymbol{\theta}}(\boldsymbol{y}) - f_{\hat{\theta}_{\eta_T,T}(\boldsymbol{x})}(\boldsymbol{y}) \right| \upsilon(\mathrm{d}\boldsymbol{y})\, \upsilon(\mathrm{d}\boldsymbol{x}) \\ &\leq 3(d_T+1)\, d_T^{1/2} \upsilon\left[\exp\left(\frac{T^{1/2}|X_1|}{4}\right)\right] \upsilon[|X_1|] \sup_{\boldsymbol{\theta}\in\Theta_T} \|\boldsymbol{\theta}\| \\ &\leq 6\log^{3/2} T\, \upsilon\left[\exp\left(\frac{T^{1/2}|X_1|}{4}\right)\right] \upsilon[|X_1|]\,. \end{aligned}$$

As X_1 is centered and normally distributed of variance γ_0, $\upsilon[|X_1|] = (2\gamma_0/\pi)^{1/2}$ and $\upsilon[\exp(T^{1/2}|X_1|/4)] = \gamma_0 T^{1/2} \exp(\gamma_0 T/32)/4$.
From $n \geq M^*(T,\varepsilon) = 9\gamma_0^3 T^2 \exp(\gamma_0 T/16)/(2\pi\varepsilon^2 \log^3 T)$ the result of Theorem 3.4.1 is reached. This bound of $M(T,\varepsilon)$ is prohibitive from a computational viewpoint. That is why we limit the number of iterations to a fixed n^*.
What we obtain from MCMC is $\bar{f}_{\eta_T,T,n}(\boldsymbol{y}|X) = \bar{\boldsymbol{\theta}}'_{\eta_T,T,n}(X)\,\boldsymbol{y}_{1:d_T}$ with $\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X) = \sum_{i=0}^{n-1} \boldsymbol{\theta}_{\eta_T,T,i}(X)/n$. Remark that $\bar{f}_{\eta_T,T,n}(\cdot|X) = f_{\bar{\theta}_{\eta_T,T,n}(X)}$. The risk is expressed as

$$R\left(\bar{f}_{\eta_T,T,n}(\cdot|X)\middle|X,\Phi_{\eta_T,T}\right) = \frac{\left(2\left(\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X)-\boldsymbol{\theta}\right)'\Gamma(Y)\left(\bar{\boldsymbol{\theta}}_{\eta_T,T,n}(X)-\boldsymbol{\theta}\right)+2\sigma^2\right)^{1/2}}{\pi^{1/2}}.$$

3.5.2 Numerical work

Consider 100 realisations of an autoregressive processes X simulated with the same $\boldsymbol{\theta} \in s_d(\delta)$ for $d = 8$ and $\delta = 3/4$ and with $\sigma = 1$. Let $\boldsymbol{c}^{(i)}$, $i = 1, 2$ the sequences defining two different priors in the model order:

1. $c_k^{(1)} = k^{-2}$, the sparsity is favoured,
2. $c_k^{(2)} = \mathrm{e}^{-k}$, the sparsity is strongly favoured.

For each sequence $\boldsymbol{c}$ and for each value of $T \in \{2^j, j = 6, \ldots, 12\}$ we compute $\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}$, the MCMC approximation of the Gibbs estimator using Algorithm 4 with $\eta = \eta_T$.

Algorithm 4: Independent Hastings Sampler

input the sample $X_{1:T}$ of X, the prior $\boldsymbol{c}$, the learning rate η, the generators $\pi_{k,T}$ for $k = 1, \ldots, d_T$ and a maximum iterations number n^*;
initialization $\boldsymbol{\theta}_{\eta,T,0} = \mathbf{0}$;
for *i=1* **to** $n^* - 1$ **do**
  generate $k \in \{1, \ldots, d_T\}$ using the prior $\boldsymbol{c}$;
  generate $\boldsymbol{\theta}_{candidate} \sim \pi_{k,T}$;
  generate $U \sim \mathcal{U}(0, 1)$;
  if $U \leq \alpha_{\eta,T,X}(\boldsymbol{\theta}_{\eta,T,i-1}, \boldsymbol{\theta}_{candidate})$ **then**
    $\boldsymbol{\theta}_{\eta,T,i} = \boldsymbol{\theta}_{candidate}$ **else**
      $\boldsymbol{\theta}_{\eta,T,i} = \boldsymbol{\theta}_{\eta,T,i-1}$;
return $\bar{\boldsymbol{\theta}}_{\eta,T,n^*}(X) = \sum_{i=0}^{n^*-1} \boldsymbol{\theta}_{\eta,T,k}(X) / n^*$.

The acceptance rate is computed as $\alpha_{\eta,T,X}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \exp(\eta r_T(\boldsymbol{\theta}_1 | X) - \eta r_T(\boldsymbol{\theta}_2 | X))$.
Algorithm 3 used by the distributions $\pi_{k,T}$ generates uniform random vectors on $s_k(1)$ by the method described in Beadle and Djurić (1999). It relies in the Levinson-Durbin recursion algorithm. We also implemented the numerical improvements of Andrieu and Doucet (1999).
Set $\varepsilon = 0.1$. Figure 3.1 displays the $(1 - \varepsilon)$-quantiles in data $R(\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}(X)) - (2/\pi)^{1/2}\sigma^2$ for $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ using different values of n^*.

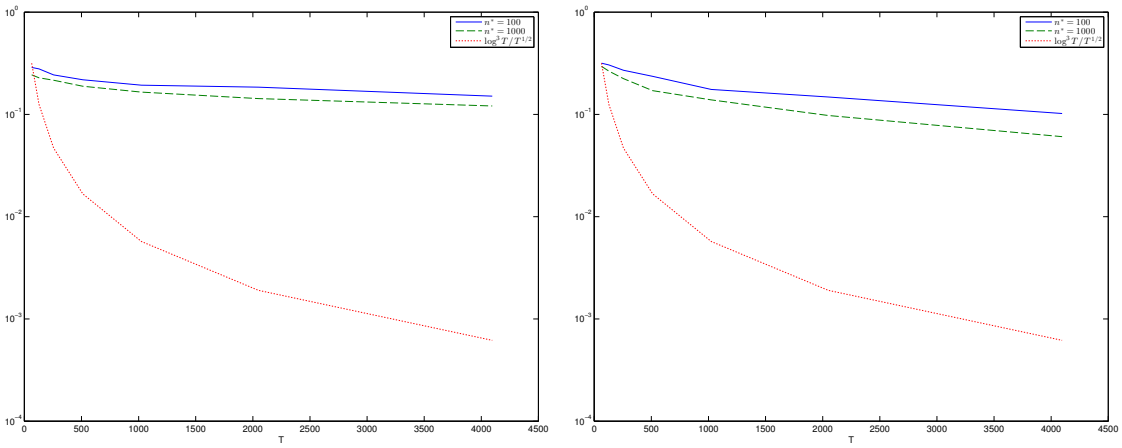


Figure 3.1 : The plots represent the 0.9-quantiles in data $R(\bar{\boldsymbol{\theta}}_{\eta_T,T,n^*}(X)) - (2/\pi)^{1/2}\sigma^2$ for $T = 32, 64, \ldots, 4096$. The graph on the left corresponds to the order prior $c_k^{(1)} = k^{-2}$ while that on the right corresponds to $c_k^{(2)} = \mathrm{e}^{-k}$. The solid curves were plotted with $n^* = 100$, the dashed ones with $n^* = 1000$ and as a reference, the dotted curve is proportional to $\log^3 T/T^{1/2}$.

Note that, for the proposed algorithm the risk in prediction decreases very slowly when the number T of observations grows and the number of MCMC iterations remains constant. If $n^* = 1000$ the decaying rate is faster than if $n^* = 100$ for smaller values of T. For $T \geq 2000$ we observe that both rates are roughly the same in the logarithmic scale. This behaviour is similar in both cases presented in Figure 3.1. As expected, the risk of the approximated predictor does not converge as $\log^3 T/T^{1/2}$.

3.6 Discussion

There are two sources of error in our method: prediction (of the exact Gibbs predictor) and approximation (using the MCMC). The first one decays when T grows and the obtained guarantees for the second one explode. We found a possibly pessimistic upper bound for $M(T, \varepsilon)$. The exponential growing of this bound is the main weakness of our procedure. The use of a better adapted proposal in the MCMC algorithm needs to be investigated. The Metropolis Langevin Algorithm (see Atchadé (2006)) gives us an insight in this direction. We refer also to the contribution of Dalalyan and Tsybakov (2012). However it is encouraging to see that, in the analysed simulation study, the risk of $\bar{f}_{\eta_T,T,n^*}(\cdot|X)$ does not increase with T.

Acknowledgements

The author is specially thankful to François Roueff, Christophe Giraud, Peter Weyer-Brown and the two referees for their extremely careful readings and highly pertinent remarks which substantially improved the paper. This work has been partially supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île de France (RDM-IdF) for the period 2012 - 2015 and by the Labex LMH (ANR-11-IDEX-003-02).

3.7 Technical proofs

3.7.1 Proof of Theorem 3.3.1

The proof of Theorem 3.3.1 is based on the same tools used by Alquier and Wintenberger (2012) up to Lemma 3. For the sake of completeness we quote the essential ones.
We denote by $\mathcal{M}_+^1(F)$ the set of probability measures on the measurable space $(F, \mathcal{F})$. Let $\rho, \nu \in \mathcal{M}_+^1(F)$, $\mathcal{K}(\rho, \nu)$ stands for the Kullback-Leibler divergence of ν from ρ.

$$\mathcal{K}(\rho,\nu) = \begin{cases} \int \log \frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\boldsymbol{\theta})\,\rho(\mathrm{d}\boldsymbol{\theta}) & \text{, if } \rho \ll \nu, \\ +\infty & \text{, otherwise .} \end{cases}$$

The first lemma can be found in (Catoni, 2004, Equation 5.2.1).

Lemma 1 (Legendre transform of the Kullback divergence function)**.** *Let $(F, \mathcal{F})$ be any measurable space. For any $\nu \in \mathcal{M}_+^1(F)$ and any measurable function $h : F \to \mathbb{R}$ such that $\nu[\exp(h)] < \infty$ we have,*

$$\nu[\exp(h)] = \exp\left(\sup_{\rho \in \mathcal{M}_+^1(F)} (\rho[h] - \mathcal{K}(\rho, \nu))\right),$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of ν, the supremum with respect to ρ in the right-hand side is reached by the Gibbs measure $\nu\{h\}$.

For a fixed $C > 0$, let $\widetilde{\xi}_t^{(C)} = \max\{\min\{\xi_t, C\}, -C\}$. Consider $\widetilde{X}_t = H(\widetilde{\xi}_t^{(C)}, \widetilde{\xi}_{t-1}^{(C)}, \ldots)$.
Denote $\widetilde{X} = (\widetilde{X}_t)_{t\in\mathbb{Z}}$ and by $\widetilde{R}(\boldsymbol{\theta})$ and $\widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right)$ the respective exact and empirical risks associated with $\widetilde{X}$ in $\boldsymbol{\theta}$.

$$\widetilde{R}(\boldsymbol{\theta}) = \mathbb{E}\left[\ell\left(\widetilde{\widetilde{X}}_t^{\boldsymbol{\theta}}, \widetilde{X}_t\right)\right],$$
$$\widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right) = \frac{1}{T - d(\boldsymbol{\theta})} \sum_{t=d(\boldsymbol{\theta})+1}^{T} \ell\left(\widetilde{\widetilde{X}}_t^{\boldsymbol{\theta}}, \widetilde{X}_t\right),$$

where $\widetilde{\widetilde{X}}_t^{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}((\widetilde{X}_{t-i})_{i\geq 1})$.
This thresholding is interesting because truncated CBS are weakly dependent processes (see (Alquier and Wintenberger, 2012, Section 4.2)).
A Hoeffding type inequality introduced in (Rio, 2000, Theorem 1) provides useful controls on the difference between empirical and exact risks of a truncated process.

Lemma 2 (Laplace transform of the risk)**.** *Let ℓ be a loss function meeting Assumption (**L**) and $X = (X_t)_{t\in\mathbb{Z}}$ a process satisfying Assumption (**M**). For all $T \geq 2$, any $\{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_T\}$ satisfying Assumption (**P**-1), Θ_T such that d_T, defined in (3.2.7), is at most $T/2$, any truncation level $C > 0$, $\eta \geq 0$ and $\boldsymbol{\theta} \in \Theta_T$ we have,*

$$\mathbb{E}\left[\exp\left(\eta\left(\widetilde{R}(\boldsymbol{\theta}) - \widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right)\right)\right)\right] \leq \exp\left(\frac{4\eta^2 k^2(T, C)}{T}\right), \tag{3.7.1}$$

and

$$\mathbb{E}\left[\exp\left(\eta\left(\widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right) - \widetilde{R}(\boldsymbol{\theta})\right)\right)\right] \leq \exp\left(\frac{4\eta^2 k^2(T, C)}{T}\right), \tag{3.7.2}$$

*where $k(T, C) = 2^{1/2} C K (1 + L_T)\left(A_* + \tilde{A}_*\right)$. The constants $\tilde{A}_*$ and A_* are defined in (3.2.4) and (3.2.5) respectively, K and L_T in Assumptions (**L**) and (**P**-1) respectively.*

The following lemma is a slight modification of (Alquier and Wintenberger, 2012, Lemma 6.5). It links the two versions of the empirical risk: original and truncated.

Lemma 3. *Suppose that Assumption (L) holds for the loss function ℓ, Assumption (M) holds for $X = (X_t)_{t\in\mathbb{Z}}$ and Assumption (I) holds for the innovations with $\zeta = A_*$; A_* is defined in (3.2.5). For all $T \geq 2$, any $\{f_\theta, \theta \in \Theta_T\}$ meeting Assumption (P-1) with Θ_T such that d_T, defined in (3.2.7), is at most $T/2$, any truncation level $C > 0$ and any $0 \leq \eta \leq T/4\,(1 + L_T)$ we have,*

$$\mathbb{E}\left[\exp\left(\eta \sup_{\theta\in\Theta_T}\left|r_T\left(\theta\,|X\right) - \widetilde{r}_T\left(\theta\,\middle|\widetilde{X}\right)\right|\right)\right] \;\leq\; \exp\left(\eta\varphi\left(T, C, \eta\right)\right) ,$$

where

$$\varphi(T, C, \eta) \;=\; 2K(1 + L_T)\phi(A_*)\left(\frac{A_* C}{\exp\left(A_* C\right) - 1} + \eta\frac{4K(1 + L_T)}{T}\right) ,$$

with K and L_T defined in Assumptions (L) and (P-1) respectively.

Finally we present a result on the aggregated predictor defined in (3.3.1). The proof is partially inspired by that of (Alquier and Wintenberger, 2012, Theorem 3.2).

Lemma 4. *Let ℓ be a loss function such that Assumption (L) holds and let $X = (X_t)_{t\in\mathbb{Z}}$ a process satisfying Assumption (M) with probability distribution υ. For each $T \geq 2$ let $\{f_\theta, \theta \in \Theta_T\}$ be a set of predictors and $\pi_T \in \mathcal{M}_+^1(\Theta_T)$ any prior probability distribution on Θ_T. We build the predictor $\hat{f}_{\eta,T}(\cdot\,|X)$ following (3.3.1) with any $\eta > 0$. For any $\varepsilon > 0$ and any truncation level $C > 0$, with υ-probability at least $1 - \varepsilon$ we have,*

$$\begin{aligned} R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\middle|\,X\right) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho\left[R\right] + \frac{2\mathcal{K}\left(\rho, \pi_T\right)}{\eta}\right\} + \frac{2\log\left(2/\varepsilon\right)}{\eta} \\ + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\widetilde{R} - \widetilde{r}_T\right)\right)\right]\right) + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\widetilde{r}_T - \widetilde{R}\right)\right)\right]\right) \\ + \frac{2}{\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\sup_{\theta\in\Theta_T}\left|r_T\left(\theta\,|X\right) - \widetilde{r}_T\left(\theta\,\middle|\widetilde{X}\right)\right|\right)\right]\right) . \end{aligned}$$

Proof. We use Tonelli's theorem and Jensen's inequality with the convex function g to obtain an upper bound for $R(\hat{f}_{\eta,T}(\cdot|X)|X)$

$$\begin{aligned} R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\middle|\,X\right) &= \int_{\mathcal{X}^{\mathbb{Z}}} g\left(\int_{\Theta_T}\left(f_\theta\left(\left(y_{t-i}\right)_{i\geq 1}\right) - y_t\right)\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left(\mathrm{d}\theta\right)\right)\upsilon\left(\mathrm{d}\mathbf{y}\right) \\ &\leq \int_{\mathcal{X}^{\mathbb{Z}}}\left[\int_{\Theta_T} g\left(f_\theta\left(\left(y_{t-i}\right)_{i\geq 1}\right) - y_t\right)\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left(\mathrm{d}\theta\right)\right]\upsilon\left(\mathrm{d}\mathbf{y}\right) \\ &= \int_{\Theta_T}\left[\int_{\mathcal{X}^{\mathbb{Z}}} g\left(f_\theta\left(\left(y_{t-i}\right)_{i\geq 1}\right) - y_t\right)\upsilon\left(\mathbf{y}\right)\right]\pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left(\mathrm{d}\theta\right) = \pi_T\left\{-\eta r_T\left(\cdot\,|X\right)\right\}\left[R\right] . \end{aligned}$$

Observe that, since the process is stationary, the previous computations are valid for any $t = 1, \ldots, T$. As explained in Section 2.3.1.1, the prediction risk, a priori defined as a mean, is the expectation of the error in just one instance.
In the remainder of this proof we search for upper bounding $\pi_T \{-\eta r_T(\cdot|X)\}[R]$.
First, we use the relationship:

$$R - r_T(\cdot|X) = \left(\widetilde{R} - \widetilde{r}_T\left(\cdot\middle|\widetilde{X}\right)\right) + \left(R - \widetilde{R}\right) - \left(r_T(\cdot|X) - \widetilde{r}_T\left(\cdot\middle|\widetilde{X}\right)\right) . \tag{3.7.3}$$

For the sake of simplicity and while it does not disrupt the clarity, we lighten the notation of r_T and $\widetilde{r}_T$. We now suppose that in the place of $\boldsymbol{\theta}$ we have a random variable distributed as $\pi_T \in \mathcal{M}_+^1(\Theta_T)$. This is taken into account in the following expectations. The identity (3.7.3) and the Cauchy-Schwarz inequality lead to

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{\eta}{2}(R - r_T)\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\eta}{2}\left(\widetilde{R} - \widetilde{r}_T\right)\right)\exp\left(\frac{\eta}{2}\left(\left(R - \widetilde{R}\right) - (r_T - \widetilde{r}_T)\right)\right)\right] \\
&\leq \left(\mathbb{E}\left[\exp\left(\eta\left(\widetilde{R} - \widetilde{r}_T\right)\right)\right]\mathbb{E}\left[\exp\left(\eta\left(\left(R - \widetilde{R}\right) - (r_T - \widetilde{r}_T)\right)\right)\right]\right)^{1/2} \\
&\leq \left(\mathbb{E}\left[\exp\left(\eta\left(\widetilde{R} - \widetilde{r}_T\right)\right)\right]\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|\left(R - \widetilde{R}\right)(\boldsymbol{\theta}) - (r_T - \widetilde{r}_T)(\boldsymbol{\theta})\right|\right)\right]\right)^{1/2} .
\end{aligned} \tag{3.7.4}$$

Observe now that $R(\boldsymbol{\theta}) = \mathbb{E}[r_T(\boldsymbol{\theta}|X)]$ and $\widetilde{R}(\boldsymbol{\theta}) = \mathbb{E}[\widetilde{r}_T(\boldsymbol{\theta}|\widetilde{X})]$. Jensen's inequality for the exponential function gives that

$$\begin{aligned}
\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|R(\boldsymbol{\theta}) - \widetilde{R}(\boldsymbol{\theta})\right|\right) &\leq \exp\left(\eta\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right)\right|\right]\right) \\
&\leq \mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right)\right|\right)\right] .
\end{aligned} \tag{3.7.5}$$

From (3.7.5) we see that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|\left(R - \widetilde{R}\right)(\boldsymbol{\theta}) - (r_T - \widetilde{r}_T)(\boldsymbol{\theta})\right|\right)\right] \\
&\quad\leq \mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|R(\boldsymbol{\theta}) - \widetilde{R}(\boldsymbol{\theta})\right|\right)\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right)\right|\right)\right] \\
&\quad\leq \left(\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right)\right|\right)\right]\right)^2 .
\end{aligned} \tag{3.7.6}$$

Combining (3.7.4) and (3.7.6) we obtain

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{\eta}{2}(R - r_T(\cdot|X))\right)\right] \leq &\left(\mathbb{E}\left[\exp\left(\eta\left(\widetilde{R} - \widetilde{r}_T\right)\right)\right]\right)^{1/2} \\
&\mathbb{E}\left[\exp\left(\eta \sup_{\boldsymbol{\theta}\in\Theta_T}\left|r_T(\boldsymbol{\theta}|X) - \widetilde{r}_T\left(\boldsymbol{\theta}\middle|\widetilde{X}\right)\right|\right)\right] .
\end{aligned} \tag{3.7.7}$$

Let $L_{\eta,T,C} = \log((\mathbb{E}[\exp(\eta(\widetilde{R} - \widetilde{r}_T))])^{1/2}\mathbb{E}[\exp(\eta \sup_{\theta\in\Theta_T} |r_T(\theta|X) - \widetilde{r}_T(\theta|\widetilde{X})|)])$. Remark that the left term of (3.7.7) is equal to the integral of the expression enclosed in brackets with respect to the measure $\upsilon \times \pi_T$. Changing η by 2η and thanks to Lemma 1 we get

$$\upsilon\left[\exp\left(\sup_{\rho\in\mathcal{M}_+^1(\Theta_T)} (\eta\rho[R - r_T(\cdot|X)] - \mathcal{K}(\rho,\pi_T))\right)\right] \leq \exp\left(L_{2\eta,T,C}\right) .$$

Markov's inequality implies that for all $\varepsilon > 0$, with υ- probability at least $1 - \varepsilon$

$$\sup_{\rho\in\mathcal{M}_+^1(\Theta_T)} (\eta\rho[R - r_T(\cdot|X)] - \mathcal{K}(\rho,\pi_T)) - \log\left(\frac{1}{\varepsilon}\right) - L_{2\eta,T,C} \leq 0 .$$

Hence, for any $\pi_T \in \mathcal{M}_+^1(\Theta_T)$ and $\eta > 0$, with υ- probability at least $1 - \varepsilon$, for all $\rho \in \mathcal{M}_+^1(\Theta_T)$

$$\rho[R - r_T(\cdot|X)] - \frac{1}{\eta}\mathcal{K}(\rho,\pi_T) - \frac{1}{\eta}\log\left(\frac{1}{\varepsilon}\right) - \frac{L_{2\eta,T,C}}{\eta} \leq 0 . \tag{3.7.8}$$

By setting $\rho = \pi_T\{-\eta r_T(\cdot|X)\}$ and relying on Lemma 1, we have

$$\begin{aligned}
\mathcal{K}(\pi_T\{-\eta r_T\},\pi_T) &= \pi_T\{-\eta r_T\}\left[\log\frac{\mathrm{d}\pi_T\{-\eta r_T\}}{\mathrm{d}\pi_T}\right] = \pi_T\{-\eta r_T\}\left[\log\frac{\exp(-\eta r_T)}{\pi_T[\exp(-\eta r_T)]}\right] \\
&= \pi_T\{-\eta r_T\}[-\eta r_T] - \log(\pi_T[\exp(-\eta r_T)]) \\
&= \pi_T\{-\eta r_T\}[-\eta r_T] + \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\{\rho[\eta r_T] + \mathcal{K}(\rho,\pi_T)\}
\end{aligned}$$

Using (3.7.8) with $\rho = \pi_T\{-\eta r_T(\cdot|X)\}$ it follows that, with υ- probability at least $1 - \varepsilon$,

$$\pi_T\{-\eta r_T(\cdot|X)\}[R] \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}\left\{\rho[r_T(\cdot|X)] + \frac{\mathcal{K}(\rho,\pi_T)}{\eta}\right\} + \frac{\log(1/\varepsilon)}{\eta} + \frac{L_{2\eta,T,C}}{\eta} .$$

To upper bound $\rho[r_T(\cdot|X)]$ we use an upper bond on $\rho[r_T(\cdot|X) - R]$. We obtain an inequality similar to (3.7.8) with $\rho[R - r_T(\cdot|X)]$ replaced by $\rho[r_T(\cdot|X) - R]$ and $L_{\eta,T,C}$ replaced by $L'_{\eta,T,C} = \log((\mathbb{E}[\exp(\eta(\widetilde{r}_T - \widetilde{R}))])^{1/2}\mathbb{E}[\exp(\eta \sup_{\theta\in\Theta_T} |r_T(\theta|X) - \widetilde{r}_T(\theta|\widetilde{X})|)])$. This provides us another inequality satisfied with υ- probability at least $1 - \varepsilon$. To obtain a υ-probability of the intersection larger than $1 - \varepsilon$ we apply previous computations with $\varepsilon/2$ instead of ε and hence,

$$\begin{aligned}
\pi_T\{-\eta r_T(\cdot|X)\}[R] \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_T)}&\left\{\rho[R] + \frac{2\mathcal{K}(\rho,\pi_T)}{\eta}\right\} + \frac{2\log(2/\varepsilon)}{\eta} \\
&+ \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\widetilde{R} - \widetilde{r}_T\right)\right)\right]\right) + \frac{1}{2\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\left(\widetilde{r}_T - \widetilde{R}\right)\right)\right]\right) \\
&+ \frac{2}{\eta}\log\left(\mathbb{E}\left[\exp\left(2\eta\sup_{\theta\in\Theta_T}\left|r_T(\theta|X) - \widetilde{r}_T\left(\theta\middle|\widetilde{X}\right)\right|\right)\right]\right) .
\end{aligned}$$

□

We can now prove Theorem 3.3.1.

Proof. Let $\pi_{0,C}$ denote the distribution on $\mathcal{X}^{\mathbb{Z}} \times \mathcal{X}^{\mathbb{Z}}$ of the couple $(X, \widetilde{X})$. Fubini's theorem and (3.7.1) of Lemma 2 imply that

$$\mathbb{E}\left[\exp\left(2\eta\left(\widetilde{R}-\widetilde{r}_T\right)\right)\right] = \pi_{0,C}\times\pi_T\left[\exp\left(2\eta\left(\widetilde{R}-\widetilde{r}_T\right)\right)\right] = \pi_T\times\pi_{0,C}\left[\exp\left(2\eta\left(\widetilde{R}-\widetilde{r}_T\right)\right)\right] \leq \exp\left(\frac{16\eta^2k^2(T,C)}{T}\right). \quad (3.7.9)$$

Using (3.7.2), we analogously get

$$\mathbb{E}\left[\exp\left(2\eta\left(\widetilde{r}_T-\widetilde{R}\right)\right)\right] \leq \exp\left(\frac{16\eta^2k^2(T,C)}{T}\right). \quad (3.7.10)$$

Consider the set of probability measures $\{\rho_{\theta_T,\Delta}, T\geq 2, 0\leq\Delta\leq\Delta_T\}\subset\mathcal{M}_+^1(\Theta_T)$, where $\boldsymbol{\theta}_T$ is the parameter defined by Assumption (P-4) and $\rho_{\theta_T,\Delta}(\boldsymbol{\theta})\propto\pi_T(\boldsymbol{\theta})\mathbb{1}_{B(\boldsymbol{\theta}_T,\Delta)\cap\Theta_T}(\boldsymbol{\theta})$. Lemma 4, together with Lemma 3, (3.7.9) and (3.7.10) guarantee that for all $0<\eta\leq T/8(1+L_T)$

$$R\left(\hat{f}_{\eta,T}(\cdot|X)\middle|X\right)\leq\inf_{0\leq\Delta\leq\Delta_T}\left\{\rho_{\theta_T,\Delta}[R]+\frac{2\mathcal{K}(\rho_{\theta_T,\Delta},\pi_T)}{\eta}\right\}+\frac{16\eta k^2(T,C)}{T}+\frac{2\log(2/\varepsilon)}{\eta}+4\varphi(T,C,2\eta). \quad (3.7.11)$$

Thanks to assumptions (L) and (P-3), for any $T\geq 2$ and $\boldsymbol{\theta}\in B(\boldsymbol{\theta}_T,\Delta)$

$$R(\boldsymbol{\theta})-R(\boldsymbol{\theta}_T)\leq K\upsilon\left[\left\|f_{\boldsymbol{\theta}}((Y_{t-i})_{i\geq 1})-f_{\boldsymbol{\theta}_T}((Y_{t-i})_{i\geq 1})\right\|\right]\leq K\mathcal{D}d_T^{1/2}\Delta. \quad (3.7.12)$$

For $T\geq 4$ Assumption (P-4) gives

$$\mathcal{K}(\rho_{\theta_T,\Delta},\pi_T)=\log\left(\frac{1}{\pi_T[B(\boldsymbol{\theta}_T,\Delta)\cap\Theta_T]}\right)\leq -n_T^{1/\gamma}\log(\Delta)-\log(C_2). \quad (3.7.13)$$

Plugging (3.7.12) and (3.7.13) into (3.7.11) and using again Assumption (P-4)

$$R\left(\hat{f}_{\eta,T}(\cdot|X)\middle|X\right)\leq R(\boldsymbol{\theta}_T)+\inf_{0\leq\Delta\leq\Delta_T}\left\{\mathcal{E}_1d_T^{1/2}\Delta-\frac{2n_T^{1/\gamma}\log(\Delta)}{\eta}\right\}+\frac{\mathcal{E}_2\eta(1+L_T)^2C^2}{T}+\frac{\mathcal{E}_3(1+L_T)C}{\exp(A_*C)-1}+\frac{2\log(2/\varepsilon)-2\log(C_2)}{\eta}+\frac{\mathcal{E}_4(1+L_T)^2\eta}{T}, \quad (3.7.14)$$

where $\mathcal{E}_1=K\mathcal{D}$, $\mathcal{E}_2=32K^2\left(A_*+\tilde{A}_*\right)^2$, $\mathcal{E}_3=8K\phi(A_*)A_*$ and $\mathcal{E}_4=32K^2\phi(A_*)$.
We upper bound d_T by $T/2$, n_T by $\log^\gamma T$ and substitute $\Delta_T=C_3/T$. Since it is difficult to minimize the right term of (3.7.14) with respect to η and C at the same time, we evaluate them in certain values to obtain a convenient upper bound.
At a fixed ε, the convergence rate of $[2\log(2/\varepsilon)-2\log(C_2)]/\eta+\mathcal{E}_4(1+L_T)^2\eta/T$ is at best $\log T/T^{1/2}$, and we get it doing $\eta\propto T^{1/2}/\log T$. As $\eta\leq T/8(1+L_T)$ we set $\eta=\eta_T=T^{1/2}/(4\log T)$.
The order of the already chosen terms is $\log^3 T/T^{1/2}$, doing $C=\log T/A_*$ we preserve it. Taking into account that $R(\boldsymbol{\theta}_T)\leq\inf_{\theta\in\Theta_T}R(\boldsymbol{\theta})+C_1\log^3 T/T^{1/2}$ the result follows. □

3.7.2 Proof of Proposition 1

Considering that Assumption (L) holds we get

$$\left|R\left(\bar{f}_{\eta,T,n}\left(\cdot\,|X\right)\middle|\,X,\Phi_{\eta,T}\right)-R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\middle|\,X,\Phi_{\eta,T}\right)\right|\leq K\int_{\mathcal{X}^{\mathbb{Z}}}\left|\bar{f}_{\eta,T,n}\left(\boldsymbol{y}\,|X\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|X\right)\right|\upsilon\left(\mathrm{d}\boldsymbol{y}\right)$$

Observe that the last expression depends on $X_{1:T}$ and $\Phi_{\eta,T}\left(X\right)$. We bound the expectation to infer a bound in probability.
Tonelli's theorem and Jensen's inequality lead to

$$\nu_{\eta,T}\left[\left|R\left(\bar{f}_{\eta,T,n}\left(\cdot\,|X\right)\middle|\,X,\Phi_{\eta,T}\right)-R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\middle|\,X,\Phi_{\eta,T}\right)\right|\right]\leq$$
$$K\int_{\mathcal{X}^{\mathbb{Z}}}\int_{\mathcal{X}^{\mathbb{Z}}}\left(\int_{\Theta_T^{\mathbb{N}}}\left|\bar{f}_{\eta,T,n}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|^2\mu_{\eta,T}\left(\mathrm{d}\boldsymbol{\phi}\,|\boldsymbol{x}\right)\right)^{1/2}\upsilon\left(\mathrm{d}\boldsymbol{y}\right)\upsilon\left(\mathrm{d}\boldsymbol{x}\right)\ . \quad (3.7.15)$$

We are then interested in upper bounding the expression under the square root. To that end, we use (Łatuszyński and Niemiro, 2011, Theorem 3.1) which implies that for any $\boldsymbol{x}$

$$\int_{\Theta_T^{\mathbb{N}}}\left|\bar{f}_{\eta,T,n}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|^2\mu_{\eta,T}\left(\mathrm{d}\boldsymbol{\phi}\,|\boldsymbol{x}\right)\leq$$
$$\sup_{\boldsymbol{\theta}\in\Theta_T}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{y}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right)^2\left(\frac{4}{\beta_{\eta,T}\left(\boldsymbol{x}\right)}-3\right)\left(\frac{1}{n}+\frac{2}{n^2\beta_{\eta,T}\left(\boldsymbol{x}\right)}\right)\ .$$

Plugging this on (3.7.15), using that $n\geq 1$ and that

$$\left(\left(4-3\beta_{\eta,T}\left(\boldsymbol{x}\right)\right)\left(2+\beta_{\eta,T}\left(\boldsymbol{x}\right)\right)\right)^{1/2}\leq 3\ ,$$

we obtain the following

$$\nu_{\eta,T}\left[\left|R\left(\bar{f}_{\eta,T,n}\left(\cdot\,|X\right)\middle|\,X,\Phi_{\eta,T}\right)-R\left(\hat{f}_{\eta,T}\left(\cdot\,|X\right)\middle|\,X,\Phi_{\eta,T}\right)\right|\right]\leq$$
$$\frac{3K}{n^{1/2}}\int_{\mathcal{X}^{\mathbb{Z}}}\frac{1}{\beta_{\eta,T}\left(\boldsymbol{x}\right)}\int_{\mathcal{X}^{\mathbb{Z}}}\sup_{\boldsymbol{\theta}\in\Theta_T}\left|f_{\boldsymbol{\theta}}\left(\boldsymbol{y}\right)-\hat{f}_{\eta,T}\left(\boldsymbol{y}\,|\boldsymbol{x}\right)\right|\upsilon\left(\mathrm{d}\boldsymbol{y}\right)\upsilon\left(\mathrm{d}\boldsymbol{x}\right)\ .$$

The result follows from Markov's inequality.

4 Aggregation of predictors for non-stationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes

Abstract

In this work, we study the problem of aggregating a finite number of predictors for non-stationary sub-linear processes. We provide oracle inequalities relying essentially on three ingredients: 1) a uniform bound of the ℓ^1 norm of the time varying sub-linear coefficients, 2) a Lipschitz assumption on the predictors and 3) moment conditions on the noise appearing in the linear representation. Two kinds of aggregations are considered giving rise to different moment conditions on the noise and more or less sharp oracle inequalities. We apply this approach for deriving an adaptive predictor for locally stationary time varying autoregressive (TVAR) processes. It is obtained by aggregating a finite number of well chosen predictors, each of them enjoying an optimal minimax convergence rate under specific smoothness conditions on the TVAR coefficients. We show that the obtained aggregated predictor achieves a minimax rate while adapting to the unknown smoothness. To prove this result, a lower bound is established for the minimax rate of the prediction risk for the TVAR process. Numerical experiments complete this study. An important feature of this approach is that the aggregated predictor can be computed recursively and is thus applicable in an online prediction context.

4.1 Introduction

In many applications where high frequency data are observed, we wish to predict the next values of this time series through an online prediction learning algorithm able to process a large amount of data. The classical stationarity assumption on the distribution of the observations has to be weakened to take into account some smooth evolution of the environment. From a statistical modelling point of view this is described by some time varying parameters. In order to sequentially track them from high-frequency data, the algorithms must require few operations and a low storage capacity to update the parameters estimation and the prediction after each new observation. The most common online methods are least mean squares (LMS), normalised least mean squares (NLMS), regularised least squares (RLS) or Kalman. All of them rely on the choice of a gradient

step, a forgetting factor, or more generally on a tuning parameter corresponding to some *a priori* knowledge on how smoothly the local statistical distribution of the data evolves along the time. To adapt automatically to this smoothness, usually unknown in practice, we propose to use an exponentially weighted aggregation of several such predictors, with various tuning parameters. We emphasize that to meet the online constraint, we cannot use methods that require a large amount of computations (such as cross validation).

The exponential weighting technique in aggregation have been developed in parallel in the machine learning community (see the seminal paper Vovk (1990)), in the statistical community (see Catoni (1997); Yang (2000a, 2004); Leung and Barron (2006), or more recently Dalalyan and Tsybakov (2008); Audibert (2009); Rigollet and Tsybakov (2012)) and in the game theory community for individual sequences prediction (see Cesa-Bianchi and Lugosi (2006) and Stoltz (2011) for recent surveys). In contrast to the classical statistical setting, in the individual sequence setting the observations are not assumed to be generated by an underlying stochastic process. The link between both settings has been analyzed in Gerchinovitz (2011) for the regression model with fixed and random designs.

Exponential weighting has also been investigated in the case of weakly dependent stationary data in Alquier and Wintenberger (2012). More recently, an approach inspired from individual sequences prediction has been studied in Anava et al. (2013) for bounded ARMA processes under some specific conditions on the (constant) ARMA coefficients.

In this contribution, we consider two possible aggregation schemes based on exponential weights which can be computed recursively. We provide oracle inequalities applying to the aggregated predictor under the following main assumptions that 1) the observations are sub-linearly with respect to a sequence of random variables with possibly time varying linear coefficients and 2) the predictors to be aggregated are Lipschitz functions of the past. An important feature of our observation model is that it embeds the well-known class of *locally stationary* processes. We refer to Dahlhaus (2009) and the references therein for a recent general view about statistical inference for locally stationary processes. As an application, we focus on a particular locally stationary model, that of the time varying autoregressive (TVAR) process. The minimax rate of certain recursive estimators of the TVAR coefficients is studied in Moulines et al. (2005). To our knowledge, there is not a well-established method on the automatic choice of the gradient step when the smoothness index is unknown. Here, we are interested in the prediction problem which is closely related to the estimation problem. We show that the proposed aggregation methods provide a solution to this question, in the sense that they give rise to recursive adaptive minimax predictors.

The paper is organized as follows. In Section 4.2, we provide oracle inequalities for the aggregated predictors under general conditions applying to non-stationary sub-linear processes. TVAR processes are introduced in Section 4.3 in a non-parametric setting based on Hölder smoothness assumptions on the TVAR coefficients. A lower bound of the prediction risk is given in this setting and this result is used to show that the proposed aggregation methods achieve the minimax adaptive rate. Section 4.4 contains the proofs of the oracle inequalities. The proof of the lower bound of the minimax prediction risk

is presented in Section 4.5. Numerical experiments illustrating these results are then described in Section 4.6. Three appendices complete this paper. Section 4.7 explains how to build non-adaptive minimax predictors which can be used in the aggregation step, Section 4.8 contains some postponed proofs and useful lemmas, and Section 4.9 provides additional results with improved aggregation rates.

4.2 ONLINE AGGREGATION OF PREDICTORS FOR NON-STATIONARY PROCESSES

4.2.1 General model

In this section, we consider a time series $(X_t)_{t\in\mathbb{Z}}$ admitting the following *non-stationary* sub-linear property with respect to the non-negative process $(Z_t)_{t\in\mathbb{Z}}$.

(**M**-1) The process $(X_t)_{t\in\mathbb{Z}}$ satisfies

$$|X_t| \le \sum_{j\in\mathbb{Z}} A_t(j)\, Z_{t-j}\ , \tag{4.2.1}$$

where $(A_t(j))_{t,j\in\mathbb{Z}}$ are non-negative coefficients such that

$$A_* := \sup_{t\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} A_t(j) < \infty\ . \tag{4.2.2}$$

Additional assumptions will be required on $(Z_t)_{t\in\mathbb{Z}}$ to deduce useful properties for $(X_t)_{t\in\mathbb{Z}}$. Note for instance that the condition on A_* in (4.2.2) guarantees that, if $(Z_t)_{t\in\mathbb{Z}}$ has a uniformly bounded L^p-norm, the convergence of the infinite sum in (4.2.1) holds almost surely and in the L^p-sense, with both convergences defining the same limit. It follows that $(X_t)_{t\in\mathbb{Z}}$ also has uniformly bounded L^p moments. Let us give some particular contexts where the representation (**M**-1) can be used.

Example 12 (Time varying linear processes). Standard weakly stationary processes such as ARMA processes (see Brockwell and Davis (2006)) admit a Wold decomposition of the form

$$X_t = \sum_{j\ge 0} a(j)\xi_{t-j}\ ,$$

where $(\xi_t)_{t\in\mathbb{Z}}$ is a weak white noise with, says, unit variance. This model, sometimes referred to as an MA(∞) representation, is often extended to a two-sided sum representation

$$X_t = \sum_{j\in\mathbb{Z}} a(j)\xi_{t-j}\ ,$$

and additional assumptions on the existence of higher moments for $(\xi_t)_{t\in\mathbb{Z}}$ or on the independence of the ξ_t's are often used for statistical inference or prediction, see

(Brockwell and Davis, 2006, Chapters 7 and 8). Because the sequence $(A_t(j))_{j\in\mathbb{Z}}$ may vary with t in (**M**-1), we may extend this standard stationary setting and also consider linear processes with time varying coefficients. In this case, we have

$$X_t = \sum_{j\in\mathbb{Z}} a_t(j)\,\xi_{t-j}\,, \tag{4.2.3}$$

where (ξ_t) is a sequence of centered independent random variables with unit variance and $(a_t(j))_{t,j\in\mathbb{Z}}$ is supposed to satisfy (4.2.2) with $A_t(j) = |a_t(j)|$, so that (**M**-1) holds with $Z_t = |\xi_t|$. For this general class of processes, statistical inference is not easily carried out : each new observation X_t comes with a new unknown sequence $(a_t(j))_{j\in\mathbb{Z}}$. However additional assumptions on this set of sequences allow to derive and study appropriate statistical inference procedures. A sensible approach in this direction is to consider a *locally stationary* model as introduced in Dahlhaus (1996b). In this framework, the set of sequences $\{(a_t(j))_{j\in\mathbb{Z}},\ 1 \le t \le T\}$ is controlled as $T \to \infty$ by artificially (but meaningfully) introducing a dependence in T, hence is written as $(a_{t,T}(j))_{j\in\mathbb{Z},1\le t\le T}$, and by approximating it with a set of sequences rescaled on the time interval $[0,1]$, $a(u,j)$, $u \in [0,1]$, $j \in \mathbb{Z}$, for example in the following way

$$\sup_{T\ge 1}\sup_{j\in\mathbb{Z}}\sum_{t=1}^{T}\left|a_{t,T}(j) - a\left(\frac{t}{T}, j\right)\right| < \infty\,.$$

Then various interesting statistical inference problems based on $X_1,\dots,X_T$ can be tackled by assuming some smoothness on the mapping $u \mapsto a(u,j)$ and, possibly, additional assumptions on the structure of the sequence $(a(u,j))_{j\in\mathbb{Z}}$ for each $u \in [0,1]$ (see Dahlhaus (2009) and the references therein).

Example 13 (TVAR model). A particular instance of Example 12 is the *time varying autoregressive* (TVAR) process, which is assumed to satisfy the recursive equation

$$X_t = \sum_{j=1}^{d}\theta_{j,t}X_{t-j} + \sigma_t\xi_t\,,$$

where $(\xi_t)_{t\in\mathbb{Z}}$ is a white noise process; see Grenier (1983). It turns out that, in the framework introduced by Dahlhaus (1996b), under suitable assumptions, such processes admit a time varying linear representation of the form (4.2.3); see Künsch (1995); Dahlhaus (1996b). In Section 4.3, we focus on such a class of processes and use the aggregation of predictors to derive adaptive minimax predictors under specific smoothness assumptions on the time varying coefficients.

Example 14 (A non-linear extension). It can also be interesting to consider non-linear extensions of Example 13. A simple example is obtained by setting

$$X_t = g_t(X_{t-1}) + \xi_t \ ,$$

where $(\xi_t)_{t\in\mathbb{Z}}$ is an i.i.d. sequence and g_t is a time varying sub-linear sequence of functions satisfying, for all t that

$$|g_t(x)| \leq \alpha|x| \ ,$$

for some $\alpha \in (0,1)$. Since g_t is no longer linear but sub-linear, such a model does not enjoy an exact linear representation of the form (4.2.3). Nevertheless, since we have

$$|X_t| \leq \alpha|X_{t-1}| + |\xi_t| \ ,$$

and iterating this equation backwards yields Assumption (**M**-1) with $Z_t = |\xi_t|$ and $A_t(j) = \alpha^j$. In the stationary case, where $g = g_t$ does not depend on t, a well-known non-linear extension is the threshold autoregressive model where g is piecewise linear; see Tong and Lim (1980).

Our goal in this section is to derive oracle bounds for the aggregation of predictors that hold for the general model (**M**-1) with one of the two following additional assumptions on $(Z_t)_{t\in\mathbb{Z}}$.

(**N**-1) The non-negative process $(Z_t)_{t\in\mathbb{Z}}$ satisfies

$$m_p := \sup_{t\in\mathbb{Z}} \mathbb{E}\left[Z_t^p\right] < \infty \ .$$

(**N**-2) The non-negative process $(Z_t)_{t\in\mathbb{Z}}$ is a sequence of independent random variables fulfilling

$$\phi(\zeta) := \sup_{t\in\mathbb{Z}} \mathbb{E}\left[\mathrm{e}^{\zeta Z_t}\right] < \infty \ .$$

Assumptions (**N**-1) and (**N**-2) appear to be quite mild. As mentioned in Example 12, basic assumptions in stationary time series usually include moments of sufficiently high order for the innovations and their independence, or rely on the Gaussian assumption, which is contained in (**N**-2). We also note that, in the context of locally stationary time series, our assumptions on the innovations are weaker than those used in the recent works Dahlhaus and Polonik (2006, 2009); Dahlhaus (2009). Precise comparisons between our assumptions and usual ones in the aggregation literature will be given after Corollary 1.

4.2.2 Aggregation of predictors

Let $(x_t)_{t\in\mathbb{Z}}$ be a real valued sequence. We say that $\widehat{x_t}$ is a predictor of x_t if it is a measurable function of $(x_s)_{s\leq t-1}$. Throughout this paper, the quality of a sequence of predictors

$(\widehat{x}_t)_{1\le t\le T}$ is evaluated for some $T \ge 1$ using the ℓ^2 loss averaged over the time period $\{1,\dots,T\}$

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{x}_t - x_t)^2 .$$

Now, given a collection of N sequences of predictors $\left\{(\widehat{x}_t^{(i)})_{1\le t\le T}, 1 \le i \le N\right\}$, we wish to sequentially derive a new predictor which predicts almost as accurately as or more accurately than the best of them.

In the present paper and for our purposes, aggregating the predictors amounts to compute a convex combination of them at each time t. This corresponds to choosing at each time t an element α_t of the simplex

$$\mathcal{S}_N = \left\{ s = (s_1,\dots,s_N) \in \mathbb{R}_+^N : \sum_{i=1}^{N} s_i = 1 \right\}. \tag{4.2.4}$$

and compute

$$\widehat{x}_t^{[\alpha_t]} = \sum_{i=1}^{N} \alpha_{i,t}\,\widehat{x}_t^{(i)} .$$

We consider two strategies of aggregation, which are studied in the context of bounded sequences in Cesa-Bianchi and Lugosi (2006); Catoni (2004). More recent contributions and extensions can be found in Gerchinovitz (2011). See also Stoltz (2011) for a pedagogical introduction. These strategies are sequential and online, meaning that

(i) to compute the aggregation weights α_t at time t, only the values of $\{\widehat{x}_s^{(i)},\ 1 \le i \le N\}$ and x_s up to time $s = t-1$ are used

(ii) the computation can be done recursively by updating a set of quantities, the number of which does not depend on t.

These two properties are met in the Algorithm 5 detailed below.

We consider in the remaining of the paper a convex aggregation of predictors

$$\widehat{x}_t = \widehat{x}_t^{[\widehat{\alpha}_t]} = \sum_{i=1}^{N} \widehat{\alpha}_{i,t}\widehat{x}_t^{(i)}, \quad 1 \le t \le T ,$$

with some specific weights $\widehat{\alpha}_{i,t}$ defined as follows.

Strategy 1: building weights from the gradient of the quadratic loss

The first strategy is to define for all $i = 1,\dots,N$ and $t = 1,\dots,T$, the weights $\widehat{\alpha}_{i,t}$ by

$$\widehat{\alpha}_{i,t} = \frac{\exp\left(-2\eta\sum_{s=1}^{t-1}\left(\sum_{j=1}^{N}\widehat{\alpha}_{j,s}\widehat{x}_s^{(j)} - x_s\right)\widehat{x}_s^{(i)}\right)}{\sum_{k=1}^{N}\exp\left(-2\eta\sum_{s=1}^{t-1}\left(\sum_{j=1}^{N}\widehat{\alpha}_{j,s}\widehat{x}_s^{(j)} - x_s\right)\widehat{x}_s^{(k)}\right)}, \tag{4.2.5}$$

with the convention that a sum over no element is zero, so $\widehat{\alpha}_{i,1} = 1/N$ for all i.
The parameter $\eta > 0$, usually called the *learning rate*, will be specified later.

Strategy 2: building weights from the quadratic loss

The second strategy is to define for all $i = 1, \dots, N$ and $t = 1, \dots, T$, the weights $\widehat{\alpha}_{i,t}$ by

$$\widehat{\alpha}_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(i)} - x_s\right)^2\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\widehat{x}_s^{(k)} - x_s\right)^2\right)}, \tag{4.2.6}$$

with again the convention that a sum over no element is zero.

Algorithm 5: Online computation of the aggregation algorithms.

parameters the learning rate η (in $(0, \infty)$) and the strategy (1 or 2);
initialization $t = 1$, $\widehat{\alpha}_t = (1/N)_{i=1,\dots,N}$;
while input *the predictions* $\widehat{x}_t^{(i)}$ *for* $i = 1, \dots, N$;
do
  $\widehat{x}_t = \widehat{x}_t^{[\widehat{\alpha}_t]} = \sum_{i=1}^{N} \widehat{\alpha}_{i,t} \widehat{x}_t^{(i)}$;
  return $\widehat{x}_t$;
  and when input *a new* x_t;
  do
    $t = t + 1$;
    for $i = 1$ **to** N **do**
      switch *strategy* **do**
        case *1*
          $v_{i,t} = \widehat{\alpha}_{i,t-1} \exp\left(-2\eta \left(\widehat{x}_{t-1}^{[\widehat{\alpha}_{t-1}]} - x_{t-1}\right) \widehat{x}_{t-1}^{(i)}\right)$;
        case *2*
          $v_{i,t} = \widehat{\alpha}_{i,t-1} \exp\left(-\eta \left(\widehat{x}_{t-1}^{(i)} - x_{t-1}\right)^2\right)$;
      $\widehat{\alpha}_t = \left(v_{i,t} / \sum_{k=1}^{N} v_{k,t}\right)_{i=1,\dots,N}$;

Both strategies yield the same algorithm up to the line where $v_{i,t}$ is computed. For sake of brevity we write only one algorithm (see Algorithm 5) and use a switch/case statement to distinguish between the two strategies. Note, however, that the choice of the strategy (1 or 2) holds for the whole sequence of predictions.
The literature (we refer to Cesa-Bianchi and Lugosi (2006)) reports that Strategy 1 provides guarantees for $(\widehat{x}_t)_{1 \le t \le T}$ compared to the best constant convex combination of predictors (the *convex regret bounds* evoked in page 94) while Strategy 2 gives them with respect to the best predictor (*best predictor regret bounds*). The regret of Strategy 1 is

of the order of $T^{-1/2}$. On the other hand, the regret of Strategy 2 is of the order of T^{-1} for a well-chosen η when both, observations and predictors are bounded. This means that depending on the context, one strategy can be more suitable that the other one. In the next section we study their application to the non-stationary sub-linear framework.

4.2.3 Oracle bounds

We establish oracle bounds on the average prediction error of the aggregated predictors. These bounds ensure that the error is equal to that associated with the best convex combination of the predictors or with the best predictor (depending on the aggregation strategy), up to two remaining terms. One remaining term depends on the number N of predictors to aggregate and the other one on the *variability* of the original process. The learning rate η can then be chosen to achieve the best trade-off between these two terms.

The second remaining term indirectly depends on the variability of the predictors. We control below this variability in terms of the variability of the original process by using the following Lipschitz property.

Definition 7. *Let $L = (L_s)_{s\geq 1}$ be a sequence of non-negative numbers. A predictor $\widehat{x}_t$ of x_t from $(x_s)_{s\leq t-1}$ is said to be L-Lipschitz if*

$$\left|\widehat{x}_t\right| \leq \sum_{s\geq 1} L_s \left|x_{t-s}\right| .$$

We more specifically consider a sequence L satisfying the following assumption.

(**L**-1) The sequence $L = (L_s)_{s\geq 1}$ satisfies

$$L_* = \sum_{j\geq 1} L_j < \infty . \tag{4.2.7}$$

This condition is trivially satisfied by constant linear predictors depending only on a finite number of previous observations, that is, $\widehat{x}_t = \sum_{s=1}^{d} L_s \; x_{t-s}$. In Section 4.7.1, we extend this case in the context of the TVAR process where the coefficients L_s are replaced by estimates of the time varying autoregressive coefficients. More generally, Assumption (**L**-1) appears to be quite natural in the general context where $\mathbb{E}[X_t|(X_{t-s})_{s\geq 1}] = f_t((X_{t-s})_{s\geq 1})$, where f_t is a Lipschitz function from $\mathbb{R}^{\mathbb{N}^*}$ to $\mathbb{R}$, with Lipschitz coefficients satisfying a condition similar to (4.2.7); see, for instance, Doukhan and Wintenberger (2008) in the case of stationary time series.
We now state two upper-bounds on the mean quadratic prediction error of the aggregated predictors defined in the previous section, when the process X fulfills the sub-linear property (**M**-1).

Theorem 4.2.1. *Assume that Assumption (**M**-1) holds. Let $\{(\widehat{X}_t^{(i)})_{1\leq t\leq T},\ 1 \leq i \leq N\}$ be a collection of sequences of L-Lipschitz predictors with L satisfying (**L**-1).*

(i) Assume that the noise Z fulfills (**N**-1) *with $p = 4$ and let $\widehat{X} = (\widehat{X}_t)_{1\le t\le T}$ denote the aggregated predictor obtained using the weights* (4.2.5) *with any $\eta > 0$. Then, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + \frac{\log N}{T\eta} + 2\eta\,(1+L_*)^4\,A_*^4 m_4\,. \quad (4.2.8)$$

(ii) Assume that the noise Z satisfies (**N**-1) *with a given $p > 2$ and let $\widehat{X} = (\widehat{X}_t)_{1\le t\le T}$ denote the aggregated predictor obtained using the weights* (4.2.6) *with any $\eta > 0$. Then, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + \frac{\log N}{T\eta} + (2\eta)^{p/2-1}A_*^p(1+L_*)^p m_p\,. \quad (4.2.9)$$

(iii) Assume that the noise Z fulfills (**N**-2) *for some positive ζ and let $\widehat{X} = (\widehat{X}_t)_{1\le t\le T}$ denote the aggregated predictor obtained using the weights* (4.2.6) *with $\eta > 0$. Then, for any*

$$\lambda \in \left(0, \frac{\zeta}{a^*(L_*+1)}\right] \quad \text{with} \quad a^* := \sup_{j\in\mathbb{Z}}\sup_{t\in\mathbb{Z}} A_t(j) \le A_*\,, \quad (4.2.10)$$

we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + \frac{\log N}{T\eta} + \frac{2}{\mathrm{e}}\lambda^{-2}\left(2+\lambda(2\eta)^{-1/2}\right)\mathrm{e}^{-\lambda(2\eta)^{-1/2}}\,(\phi(\zeta))^{\lambda A_*(1+L_*)/\zeta}\,. \quad (4.2.11)$$

Remark 1. The bounds (4.2.8), (4.2.9) and (4.2.11) are explicit in the sense that all the constants appearing in them are directly derived from those appearing in Assumptions (**M**-1), (**L**-1), (**N**-1) and (**N**-2).

The proof can be found in Section 4.4.2. A crucial ingredient of it is a lemma gathering useful adaptations of well-known inequalities applying to the aggregation of deterministic predicting sequences.

Lemma 5. *Let $(x_t)_{1\le t\le T}$ be a real valued sequence and $\{(\widehat{x}_t^{(i)})_{1\le t\le T}, 1 \le i \le N\}$ be a collection of predicting sequences. Define $(\widehat{x}_t)_{1\le t\le T}$ as the sequence of aggregated*

predictors obtained from this collection with the weights (4.2.5). Then, for any $\eta > 0$, we have

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{x}_t - x_t)^2 \le \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{x}_t^{[\nu]} - x_t\right)^2 + \frac{\log N}{T\eta} + \frac{2\eta}{T}\sum_{t=1}^{T} y_t^4\,, \tag{4.2.12}$$

where $y_t = |x_t| + \max_{1\le i\le N}|\widehat{x}_t^{(i)}|$.
Define now $(\widehat{x}_t)_{1\le t\le T}$ as the sequence of aggregated predictors obtained with the weights (4.2.6). Then, for any $\eta > 0$, we have

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{x}_t - x_t)^2 \;\le\; \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{x}_t^{(i)} - x_t\right)^2 + \frac{\log N}{T\eta} + \frac{1}{T}\sum_{t=1}^{T}\left(y_t^2 - \frac{1}{2\eta}\right)_+\,, \tag{4.2.13}$$

where $y_t = |x_t| + \max_{1\le i\le N}|\widehat{x}_t^{(i)}|$.

The proof is postponed to Section 4.4.1.
The following corollary is obtained by choosing η (and λ in Case (iii)) adequately in the three cases of Theorem 4.2.1.

Corollary 1. *Assume that Assumption (M-1) holds. Let $\{(\widehat{X}_t^{(i)})_{1\le t\le T},\ 1 \le i \le N\}$ be a collection of sequences of L-Lipschitz predictors with L satisfying (L-1).*

(i) Assume that the noise Z fulfills (N-1) with $p = 4$ and let $\widehat{X} = (\widehat{X}_t)_{1\le t\le T}$ denote the aggregated predictor obtained using the weights (4.2.5) with

$$\eta = \frac{1}{(2m_4)^{1/2}\,(1+L_*)^2\,A_*^2}\left(\frac{\log N}{T}\right)^{1/2}. \tag{4.2.14}$$

This gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + C_1\left(\frac{\log N}{T}\right)^{1/2}, \tag{4.2.15}$$

with $C_1 = 2\,(2m_4)^{1/2}\,(1+L_*)^2\,A_*^2$.

(ii) Assume that the noise Z satisfies (N-1) with a given $p > 2$ and let $\widehat{X} = (\widehat{X}_t)_{1\le t\le T}$ denote the aggregated predictor obtained using the weights (4.2.6) with

$$\eta = \frac{1}{2m_p^{2/p}(1+L_*)^2\,A_*^2}\left(\frac{\log N}{T}\right)^{2/p}. \tag{4.2.16}$$

We then have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + C_2\left(\frac{\log N}{T}\right)^{1-2/p}, \tag{4.2.17}$$

with $C_2 = 3m_p^{2/p}(1+L_*)^2\,A_*^2$.

(iii) Assume that the noise Z fulfills (**N**-2) *for some positive ζ and let $\widehat{X} = (\widehat{X}_t)_{1\le t\le T}$ denote the aggregated predictor obtained using the weights* (4.2.6) *with*

$$\eta = \frac{\zeta^2}{2(1+L_*)^2 A_*^2}\left(\log\left(\frac{T}{\log N}\right)\right)^{-2} . \tag{4.2.18}$$

Then, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + \frac{2A_*^2(L_*+1)^2}{\zeta^2}\,\frac{\log N}{T}\left\{\left(\log\left(\frac{T}{\log N}\right)\right)^2 + \frac{\phi(\zeta)}{\mathrm{e}}\left(2+\log\left(\frac{T}{\log N}\right)\right)\right\} . \tag{4.2.19}$$

(Note that when $(\log N)/T \to 0$, the term between curly brackets is equivalent to $(\log(T/\log N))^2$).

Cases (i) and (ii) in Corollary 1 follow directly from Theorem 4.2.1. Case (iii) is more delicate since it requires optimizing λ as well as η in the second line of (4.2.11). The details are postponed to Section 4.4.3.

Remark 2. We observe that the bound in (4.2.19) improves that in (4.2.17) for any $p > 2$. For $p > 4$, the remaining term $(\log N/T)^{1-2/p}$ in (4.2.17) is smaller than the remaining term $(\log N/T)^{1/2}$ in (4.2.15). Similarly, the remaining term $\log N\,(\log T)^2/T$ in (4.2.19) is smaller than $(\log N/T)^{1/2}$ in (4.2.15). Yet, we emphasize that the oracle inequalities (4.2.17) and (4.2.19) compare the prediction risk of $\widehat{X}$ to the prediction risk of the *best predictor* $\widehat{X}^{(i)}$, while the oracle inequality (4.2.15) compare the prediction risk of $\widehat{X}$ to the prediction risk of the *best convex combination of the predictors* $\widehat{X}^{(i)}$, so they cannot be directly compared.

Remark 3. As explained in Section 4.9, under the hypotheses of Cases (ii) and (iii) and for certain values of T and N, using a more involved aggregation step, we can get a new predictor satisfying an oracle inequality better than that in (4.2.15). For example, under the hypotheses of Case (iii), for $T > N^2(\log T)^6$, the remaining term $(\log N/T)^{1/2}$ in (4.2.15) can be replaced by $N(\log T)^3/T$ which is smaller; see the inequality (4.9.7) page 125. Yet, this aggregation has a prohibitive computational cost and seems difficult to implement in practice.

Remark 4. In Cases (ii) and (iii), which correspond to the weights (4.2.6), the choice of the optimal η depends on the assumptions on the noise, namely (**N**-1) or (**N**-2). Under a moment condition of order p, the optimal η is of order $(\log N/T)^{2/p}$ and under an exponential condition, it is of order $(\log T)^{-2}$. It is known from (Catoni, 2004, Proposition 2.2.1) and (Yang, 2004, Theorem 5) that η can be chosen as a constant (provided that it is small enough) under a bounded noise condition, or under an exponential moment condition on the noise for predictors at a bounded distance from the conditional mean. Hence, coarsely speaking, the heavier the tail of the noise, the smallest

η should be chosen. Observing that η allows us to tune the influence of the empirical risk on the weights from no influence at all ($\eta = 0$ yielding uniform weights) to the selection of the empirical risk minimizer ($\eta \to \infty$), the specific choices of η can be interpreted as follows : the heavier the tail of the noise, the less we can trust the empirical risk.

Remark 5. The main limitation of the result is that, in each case, the proposed optimal learning rate η relies on a precise knowledge about the process (through the values of L_*, A_*, m_p and ζ). In several aggregation algorithms (polynomial weights or exponential weights for example) the learning rate η changes on time and an η_t, calibrated from the observed data, is defined for all $t \geq 1$. These methods also allow to obtain satisfying regret bounds. As an example, under the assumptions of Case (i), (Cesa-Bianchi et al., 2005, Theorem 6) ensures that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + C\, m_4^{1/2}\left(\frac{\log N}{T}\right)^{1/2},$$

where C is an absolute constant and $\widehat{X}_t$ is built without any previous information about $(X_t)_{1\leq t\leq T}$ or its predictors. Algorithms in this fashion should be explored in all the cases of our framework.

Comparison with previous works : In the literature, prediction risk bounds of the form (4.2.15) (Case (i) of Corollary 1) are sometimes called *convex regret bounds*, and prediction risk bounds of the form (4.2.17) and (4.2.19) (Cases (ii) and (iii) of Corollary 1) are sometimes called *best predictor regret bounds*.

Sancetta (2010) exhibits convex regret bounds in a setting close to ours, namely for an online aggregation of predictors for a sequence of possibly dependent random variables. Under our moment condition (**N**-1) with $p = 4$, (Sancetta, 2010, Theorem 2) provides an upper bound similar to (4.2.15) but with our remaining term $(\log N/T)^{1/2}$ replaced by $(N\log(N)/T)^{1/2}$. Under the exponential condition (**N**-2), (Sancetta, 2010, Theorem 1) provides an upper bound similar to (4.2.15) but with a remaining term $(\log N/T)^{1/2} \times (\log(NT))^2$, which is still larger than our remaining term under moment conditions.

Best predictor regret bounds can be found in Yang (2004) for some sequences of possibly dependent random variables. The predictors are assumed to remain at a bounded distance to the conditional means and the scaled innovation noise is assumed to have either a known distribution (satisfying a certain technical condition) or an exponential moment. The regret bounds are presented in a slightly different fashion from ours but it is easy to see that a similar result as our bound (4.2.19) is obtained in this setting. However, we do not require to have bounded prediction errors and our conditions on the noise are milder.

The i.i.d. setting has received much more attention and, even if the setting is quite different, it is interesting to briefly compare our results to previous works in this case. Let us start with the convex regret bound in Case (i) of Corollary 1. Most of the existing results (see, for instance, Juditsky and Nemirovski (2000); Yang (2000a); Tsybakov (2003) or Wang et al. (2014) for recent extensions to ℓ^q aggregation) assume the predictors to be bounded and various conditions on the noise are considered (very often the noise is assumed to be Gaussian). In such settings, the best possible remaining term typically

takes the form $(\log N/T)^{1/2}$ when N is much larger than $T^{1/2}$ and of the form N/T if N is smaller than $T^{1/2}$; see (Juditsky and Nemirovski, 2000, Theorem 3.1), (Yang, 2004, Theorem 6) and (Tsybakov, 2003, Theorem 2). Hence our bound (4.2.15) is similar only in the case where N is much larger than $T^{1/2}$. However, as explained in Remark 3 and Section 4.9, when T is larger than N^2 and under the moment condition (**N**-2), we can get via a more involved aggregation procedure, a convex regret bound with a remaining term of the same order N/T up to a $(\log T)^3$ factor (see inequality (4.9.7) page 125). Let us now compare our bound (4.2.17) in Case (ii) to optimal bounds in the i.i.d. setting under moment conditions on the noise. Corollary 7.2 and Theorem 8.6 in Audibert (2009) shows that the optimal aggregation rate is $(\log N/T)^{1-2/(p+2)}$ in the i.i.d. setting with bounded predictors and moment conditions of order p on the noise. Our remaining term $(\log N/T)^{1-2/p}$ in (4.2.17) is slightly larger, yet an inspection of the proof of (Audibert, 2009, Corollary 7.2) shows that the aggregation rate would also be $(\log N/T)^{1-2/p}$ in this corollary, if the predictors were assumed to have a moment condition of order p instead of being uniformly bounded (we are not aware of any lower bound in this setting matching this rate). Finally, when the data and the predictors are bounded, the best aggregation rate is known to be $(\log N)/T$ in the i.i.d. setting; see, for example, (Audibert, 2009, Theorem 8.4). Our bound (4.2.19) in Case (iii) achieves the same rate up to a $(\log T)^2$ factor.

Last but not least, Lemma 5 extends classical results in individual sequences prediction. When observations and predictors are bounded, a well-chosen η allows to obtain *best predictor regret bounds* of the order of $(\log N)/(T\eta)$ (we refer to Haussler et al. (1998); Vovk (1998) and (Cesa-Bianchi and Lugosi, 2006, Section 3.5)). In contrast, our result does not require any bound.

4.3 TIME-VARYING AUTOREGRESSIVE (TVAR) MODEL

4.3.1 Non-parametric TVAR model

4.3.1.1 Vector norms and Hölder smoothness norms

We introduce some preliminary notation before defining the model. In the remainder of this article, vectors are denoted using boldface symbols and $|\mathbf{x}|$ denotes the Euclidean norm of $\mathbf{x}$, $|\mathbf{x}| = (\sum_i |x_i|^2)^{1/2}$.

For $\beta \in (0, 1]$ and an interval $I \subseteq \mathbb{R}$, the $\beta-$Hölder semi-norm of a function $\mathbf{f} : I \to \mathbb{R}^d$ is defined by

$$|\mathbf{f}|_\beta = \sup_{0<|s-s'|<1} \frac{|\mathbf{f}(s) - \mathbf{f}(s')|}{|s - s'|^\beta} .$$

This semi-norm is extended to any $\beta > 0$ as follows. Let $k \in \mathbb{N}$ and $\alpha \in (0, 1]$ be such that $\beta = k + \alpha$. If $\mathbf{f}$ is k times differentiable on I, we define

$$|\mathbf{f}|_\beta = |\mathbf{f}^{(k)}|_\alpha ,$$

and $|\mathbf{f}|_\beta = \infty$ otherwise. We consider the case $I = (-\infty, 1]$. For $R > 0$ and $\beta > 0$, the $(\beta, R)-$ Hölder ball is denoted by

$$\Lambda_d(\beta, R) = \left\{\mathbf{f} : (-\infty, 1] \to \mathbb{R}^d, \text{ such that } |\mathbf{f}|_\beta \leq R\right\} .$$

4.3.1.2 TVAR parameters in rescaled time

The idea of using a rescaled time with the sample size T for the TVAR parameters goes back to Dahlhaus (1996b). Since then, it has always been a central example of locally stationary linear processes. In this setting, the time varying autoregressive coefficients and variance which generate the observations $X_{t,T}$ for $1 \leq t \leq T$ are represented by functions from $[0, 1]$ to $\mathbb{R}^d$ and from $[0, 1]$ to $\mathbb{R}_+$, respectively. The definition sets of these functions are extended to $(-\infty, 1]$ in the following definition.

Definition 8 (TVAR model)**.** *Let $d \geq 1$. Let $\theta_1, \ldots, \theta_d$ and σ be functions defined on $(-\infty, 1]$ and $(\xi_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. random variables with zero mean and unit variance. For any $T \geq 1$, we say that $(X_{t,T})_{t\leq T}$ is a TVAR process with time varying parameters $\theta_1, \ldots, \theta_d, \sigma^2$ sampled at frequency T^{-1} and normalized innovations (ξ_t) if the two following assertions hold.*

(i) The process X fulfills the time varying autoregressive equation

$$X_{t,T} = \sum_{j=1}^{d} \theta_j\left(\frac{t-1}{T}\right) X_{t-j,T} + \sigma\left(\frac{t}{T}\right)\xi_t \quad \text{for } -\infty < t \leq T . \tag{4.3.1}$$

(ii) The sequence $(X_{t,T})_{t\leq T}$ is bounded in probability,

$$\lim_{M\to\infty} \sup_{-\infty<t\leq T} \mathbb{P}(|X_{t,T}| > M) = 0 .$$

This definition extends the usual definition of TVAR processes, where the time varying parameters $\theta_1, \ldots, \theta_d$ and σ^2 are assumed to be constant on $\mathbb{R}_-$; see, for example (Dahlhaus, 1996b, Page 144). The TVAR model is generally used for the sample $(X_{t,T})_{1\leq t\leq T}$. The definition of the process for negative times t can be seen as a way to define initial conditions for $X_{1-d,T}, \ldots, X_{0,T}$, which are then sufficient to compute $(X_{t,T})_{1\leq t\leq T}$ by iterating (4.3.1). However, in the context of prediction, it can be useful to consider predictors $\widehat{X}_{t,T}$ which may rely on historical data $X_{s,T}$ arbitrarily far away in the past, that is, with s tending to $-\infty$. To cope with this situation, our definition of the TVAR process $(X_{t,T})$ holds for all time indices $-\infty < t \leq T$ and we use the following definition for predictors.

Definition 9 (Predictor)**.** *For all $1 \leq t \leq T$, we say that $\widehat{X}_{t,T}$ is a predictor of $X_{t,T}$ if it is $\mathcal{F}_{t-1,T}$-measurable, where*

$$\mathcal{F}_{t,T} = \sigma\left(X_{s,T},\ s = t, t-1, t-2, \ldots\right) \tag{4.3.2}$$

is the σ-field generated by $(X_{s,T})_{s\leq t}$. For any $T \geq 1$, we denote by $\mathcal{P}_T$ the set of sequences $\widehat{X}_T = (\widehat{X}_{t,T})_{1\leq t\leq T}$ of predictors for $(X_{t,T})_{1\leq t\leq T}$, that is, the set of all processes $\widehat{X}_T = (\widehat{X}_{t,T})_{1\leq t\leq T}$ adapted to the filtration $(\mathcal{F}_{t-1,T})_{1\leq t\leq T}$.

In this general framework, the time $t = 1$ corresponds to the beginning of the aggregation procedure. Such a framework applies in two practical situations. In the first one, we start collecting data $X_{t,T}$ at $t \geq 1$ and compute several predictors $\widehat{X}^{(j)}_{t,T}$, $j = 1, \ldots, N$ from them. Thus, the resulting aggregated predictor only depends on $(X_{s,T})_{1\leq s\leq t-1}$. A somewhat different situation is when historical data is available beforehand the aggregation step, so that a given predictor $\widehat{X}^{(j)}_{t,T}$ is allowed to depend also on data $X_{s,T}$ with $s \leq 0$, while the aggregation step only starts at $t \geq 1$, and thus depends on the data $(X_{s,T})_{s\leq 0}$ only through the predictors. It is important to note that, in contrast to the usual stationary situation, having observed the process $X_{s,T}$ for infinitely many s's in the past (for all $s \leq t-1$) is not so decisive for deriving a predictor of $X_{t,T}$, since observations far away in the past may have a completely different statistical behavior.

4

4.3.1.3 Stability conditions

The next proposition proves that under standard stability conditions on the time varying parameters $\theta_1, \ldots \theta_d$ and σ^2, Condition (ii) in Definition 8 ensures the existence and uniqueness of the solution of Equation (4.3.1) for $t \leq 0$ (and thus for all $t \leq T$). We define the time varying autoregressive polynomial by

$$\boldsymbol{\theta}(z; u) \quad = \quad 1 - \sum_{j=1}^{d} \theta_j(u) z^j \ .$$

Let us denote, for any $\delta > 0$,

$$s_d(\delta) = \left\{ \boldsymbol{\theta} : (-\infty, 1] \to \mathbb{R}^d, \boldsymbol{\theta}(z; u) \neq 0, \forall |z| < \delta^{-1}, u \in [0, 1] \right\} \ . \tag{4.3.3}$$

Define, for $\beta > 0$, $R > 0$, $\delta \in (0, 1)$, $\rho \in [0, 1]$ and $\sigma_+ > 0$, the class of parameters

$$C(\beta, R, \delta, \rho, \sigma_+) = \left\{ (\boldsymbol{\theta}, \sigma) : (-\infty, 1] \to \mathbb{R}^d \times [\rho\sigma_+, \sigma_+] \ : \ \boldsymbol{\theta} \in \Lambda_d(\beta, R) \cap s_d(\delta) \right\} \ .$$

The definition of the class C is very similar to that of Moulines et al. (2005). The domain of definition in their case is $[0, 1]$ whereas it is $(-\infty, 1]$ in ours. We have the following stability result.

Proposition 2. *Assume that the time varying AR coefficients $\theta_1, \ldots, \theta_d$ are uniformly continuous on $(-\infty, 1]$ and the time varying variance σ^2 is bounded on $(-\infty, 1]$. Assume moreover that there exists $\delta \in (0, 1)$ such that $\boldsymbol{\theta} \in s_d(\delta)$. Then, there exists $T_0 \geq 1$ such that, for all $T \geq T_0$, there exists a unique process $(X_{t,T})_{t\leq T}$ which satisfies (i) and (ii) in Definition 8. This solution admits the linear representation*

$$X_{t,T} = \sum_{j=0}^{\infty} a_{t,T}(j)\, \sigma\left(\frac{t-j}{T}\right) \xi_{t-j}, \quad -\infty < t \leq T \ , \tag{4.3.4}$$

where the coefficients $(a_{t,T}(j))_{t\le T, j\ge 0}$ *satisfy that for any* $\delta_1 \in (\delta, 1)$,

$$\bar{K} = \sup_{T\ge T_0} \sup_{-\infty<t\le T} \sup_{j\ge 0} \delta_1^{-j} |a_{t,T}(j)| < \infty \, .$$

Moreover, if $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, 0, \sigma_+)$ *for some positive constants* β, R *and* σ_+, *then the constants* T_0 *and* $\bar{K}$ *can be chosen only depending on* δ_1, δ, β, *and* R.

A proof of Proposition 2 is provided in Section 4.8. This kind of result is classical under various smoothness assumptions on the parameters and initial conditions for $X_{1-k,T}$, $k = 1, \dots, d$. For instance, in Dahlhaus and Polonik (2009), bounded variations and a constant $\boldsymbol{\theta}$ for negative times are used for the smoothness assumption on $\boldsymbol{\theta}$ and for defining the initial conditions. The linear representation (4.3.4) of TVAR processes was first obtained in the seminal papers Künsch (1995); Dahlhaus (1996b). We note that an important consequence of Proposition 2 is that for any $T \ge T_0$, the process $(X_{t,T})_{t\le T}$ satisfies Assumption (**M**-1) with $Z_t = |\xi_t|$ and $A_t(j) = \left|a_{t,T}(j)\, \sigma\left((t-j)/T\right)\right|$ for $j \ge 0$. Moreover, the constant A_* in (4.2.2) is bounded independently of T, and we have, for all $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, 0, \sigma_+)$,

$$A_* \le \frac{\bar{K}\sigma_+}{1-\delta_1} \, , \tag{4.3.5}$$

where $\bar{K} > 0$ and $\delta_1 \in (0, 1)$ can be chosen only depending on δ, β and R.

4.3.1.4 Main assumptions

Based on Proposition 2, given an i.i.d. sequence $(\xi_t)_{t\in\mathbb{Z}}$ and constants $\delta \in (0, 1)$, $\rho \in [0, 1]$, $\sigma_+ > 0$, $\beta > 0$ and $R > 0$, we consider the following assumption.

(**M**-2) The sequence $(X_{t,T})_{t\le T}$ is a TVAR process with time varying standard deviation σ, time varying AR coefficients $\theta_1, \dots, \theta_d$ and innovations $(\xi_t)_{t\in\mathbb{Z}}$, and $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, \rho, \sigma_+)$.

Let ξ denote a generic random variable with the same distribution as the ξ_t's. Under Assumption (**M**-2), the distribution of $(X_{t,T})_{1-d\le t\le T}$ only depends on that of ξ and on the functions $\boldsymbol{\theta}$ and σ. For a given distribution ψ on $\mathbb{R}$ for ξ, we denote by $\mathbb{P}^{\psi}_{(\boldsymbol{\theta},\sigma)}$ the probability distribution of the whole sequence $(X_{t,T})_{t\le T}$ and by $\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}$ its corresponding expectation. The next two assumptions on the innovations are useful to prove upper bounds of the prediction error.

(**I**-1) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ satisfy $m_p := \mathbb{E}\left[|\xi|^p\right] < \infty$.

(**I**-2) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ satisfy $\phi(\zeta) := \mathbb{E}\left[\mathrm{e}^{\zeta|\xi|}\right] < \infty$.

The following one will be used to obtain a lower bound.

(I-3) The innovations $(\xi_t)_{t\in\mathbb{Z}}$ admit a density f such that

$$\kappa = \sup_{v\neq 0} v^{-2} \int f(u) \log \frac{f(u)}{f(u+v)}\, \mathrm{d}u < \infty .$$

Assumption (I-3) is standard for proving lower bounds in non-parametric regression estimation, see (Tsybakov, 2009, Chapter 2). It is satisfied by the Gaussian density with $\kappa = 1$.

4.3.1.5 Non-parametric setting

The setting of Definition 8 and of Assumptions derived thereafter is essentially non-parametric, since for given initial distribution ψ, the distribution of the observations $X_{1,T}, \ldots, X_{T,T}$ are determined by the unknown parameter function $(\boldsymbol{\theta}, \sigma)$. The doubly indexed $X_{t,T}$ refers to the fact that this distribution cannot be seen as a distribution on $\mathbb{R}^{\mathbb{Z}}$ marginalized on $\mathbb{R}^T$ as the usual time series setting but rather as a sequence of distributions on $\mathbb{R}^T$ indexed by T. It corresponds to the usual non-parametric approach for studying statistical inference based on this model. In this contribution, we focus on the prediction problem, which is to answer the question: for given smoothness conditions on $(\boldsymbol{\theta}, \sigma)$, what is the mean prediction error for predicting $X_{t,T}$ from its past? The standard non-parametric approach is to answer this question in a minimax sense by determining, for a given sequence of predictors $\widehat{X}_T = (\widehat{X}_{t,T})_{1\leq t\leq T}$, the maximal risk

$$S_T\left(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+\right) = \sup_{(\boldsymbol{\theta},\sigma)} \frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)} \left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] - \sigma^2 \left(\frac{t}{T} \right) \right) , \tag{4.3.6}$$

where

(a) $\widehat{X}_T$ is assumed to belong to $\mathcal{P}_T$ as in Definition 9,

(b) the sup is taken over $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, \rho, \sigma_+)$ within a smoothness class of functions,

(c) the expectation $\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}$ is that associated with Assumption (M-2).

The reason for subtracting the average $\sigma^2(t/T)$ over all $1 \leq t \leq T$ in this prediction risk is that it corresponds to the best prediction risk, would the parameters $(\boldsymbol{\theta}, \sigma)$ be exactly known. We observe that dividing $X_{t,T}$ by the class parameter σ_+ amounts to take $\sigma_+ = 1$. In addition, we have

$$S_T\left(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+\right) = \sigma_+^2\, S_T\left(\widehat{X}_T \sigma_+^{-1}; \psi, \beta, R, \delta, \rho, 1\right) ,$$

so the prediction problem in the class $C(\beta, R, \delta, \rho, \sigma_+)$ can be reduced to the prediction problem in the class $C(\beta, R, \delta, \rho, 1)$. Accordingly, we define the reduced minimax risk by

$$\begin{aligned} \overline{M}_T(\psi, \beta, R, \delta, \rho) &= \inf_{\widehat{X}_T \in \mathcal{P}_T} S_T\left(\widehat{X}_T; \psi, \beta, R, \delta, \rho, 1\right) \\ &= \inf_{\widehat{X}_T \in \mathcal{P}_T} \sigma_+^{-2}\, S_T\left(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+\right) \quad \text{for all } \sigma_+ > 0 . \end{aligned} \tag{4.3.7}$$

In Section 4.3.2, we provide a lower bound of the minimax rate in the case where the smoothness class is of the form $C(\beta, R, \delta, \rho, \sigma_+)$. Then, in Section 4.3.3, relying on the aggregation oracle bounds of Section 4.2.3, we derive an upper bound with the same rate as the lower bound using the same smoothness class of the parameters. Moreover, we exhibit an online predictor which does not require any knowledge about the smoothness class and which is thus minimax adaptive. In other words, it is able to adapt to the unknown smoothness of the parameters from the data. To our knowledge, such theoretical results are new for locally stationary models.

4.3.2 Lower bound

A lower bound on the minimax rate for the estimation error of $\boldsymbol{\theta}$ is given by (Moulines et al., 2005, Theorem 4). Clearly, a predictor

$$\widehat{X}_{t,T} = \sum_{k=1}^{d} \widehat{\boldsymbol{\theta}}_{t,T}(k) X_{t-k,T}$$

can be defined from an estimator $\widehat{\boldsymbol{\theta}}_{t,T}$, and the resulting prediction rate can be controlled using the estimation rate (see Section 4.7.1 for the details). The next theorem provides a lower bound of the minimax rate of the risk of *any* predictor of the process $(X_{t,T})_{1\le t\le T}$. Combining this result with Lemma 12 in the Section 4.7.1, we show that a predictor obtained by Equation (4.7.1) from a minimax rate estimator of $\boldsymbol{\theta}$ automatically achieves the minimax prediction rate.

Theorem 4.3.1. *Let $\delta \in (0,1)$, $\beta > 0$, $R > 0$ and $\rho \in [0,1]$. Suppose that Assumption (M-2) holds and assume (I-3) on the distribution ψ of the innovations. Then, we have*

$$\liminf_{T\to\infty} T^{2\beta/(1+2\beta)}\, \overline{M}_T(\psi, \beta, R, \delta, \rho) > 0\,, \tag{4.3.8}$$

where $\overline{M}_T$ is defined in (4.3.7).

The proof is postponed to Section 4.5.

4.3.3 Minimax adaptive forecasting of the TVAR process

In Arkoun (2011), an adaptive estimator of the autoregressive function of a Gaussian TVAR process of order 1 is studied. It relies on the Lepskiĭ's procedure (see Lepskiĭ (1990)), which seems difficult to implement in an online context.
Our minimax adaptive predictor is based on the aggregation of sufficiently many predictors, assuming that at least one of them converges at the minimax rate. The oracle bounds found in Section 4.2.3 imply that the aggregated predictor is minimax rate adaptive under appropriate assumptions. Seminal works using the aggregation to adapt to the minimax convergence rate are Yang (2000a) (non-parametric regression) and Yang (2000b) (density estimation); see also Catoni (2004) for a more general presentation.

In the TVAR model (M-2), it is natural to consider L-Lipschitz predictors $(\widehat{X}_{t,T})_{1\le t\le T}$ of $(X_{t,T})_{1\le t\le T}$ with a sequence L supported on $\{1,\dots,d\}$. Then L^* in (4.2.7) corresponds to the maximal ℓ^1-norm of the TVAR parameters. Since for the process itself to be stable, this norm has to be bounded independently of T, Condition (L-1) is a quite natural assumption for the TVAR model, see Section 4.7.1 for the details.

A practical advantage of the proposed procedures is that, given a set of predictors that behaves well under specific smoothness assumptions, we obtain an aggregated predictor which performs almost as well as or better than the best of these predictors, hence which behaves well without any prior knowledge on the smoothness of the unknown parameter. Such an adaptive property can be formally demonstrated by exhibiting an adaptive minimax rate for the aggregated predictor which coincides with the lower bound given in Theorem 4.3.1.

The first ingredient that we need is the following.

Definition 10 ((ψ,β)-minimax-rate predictor)**.** *Let ψ be a distribution on $\mathbb{R}$ and $\beta>0$. We say that $\widehat{X}=(\widehat{X}_T)_{T\ge1}$ is a (ψ,β)-minimax-rate sequence of predictors if, for all $T\ge1$, $\widehat{X}_T\in\mathcal{P}_T$ and, for all $\delta\in(0,1)$, $R>0$, $\rho\in(0,1]$ and $\sigma_+>0$,*

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} S_T\left(\widehat{X}_T;\psi,\beta,R,\delta,\rho,\sigma_+\right)<\infty\,, \tag{4.3.9}$$

where S_T is defined by (4.3.6).

The term *minimax-rate* in this definition refers to the fact that the maximal rate in (4.3.9) is equal to the minimax lower bound (4.3.8) for the class $C(\beta,R,\delta,\rho,\sigma_+)$. We explain in Section 4.7 how to build such predictors which are moreover L-Lipschitz for some L only depending on d. To adapt to an unknown smoothness, we rely on a collection of (ψ,β)-minimax-rate predictors with β within $(0,\beta_0)$, where β_0 is the (possibly infinite) maximal smoothness index.

Definition 11 (Locally bounded set of ψ-minimax-rate predictors)**.** *Let ψ be a distribution on $\mathbb{R}$ and $\beta_0\in(0,\infty]$. We say that $\{\widehat{X}^{(\beta)},\beta\in(0,\beta_0)\}$ is a locally bounded set of ψ-minimax-rate predictors if for each β, $\widehat{X}^{(\beta)}$ is a (ψ,β)-minimax-rate predictor and if moreover, for all $\delta\in(0,1)$, $R>0$, $\rho\in(0,1]$, $\sigma_+>0$ and for each closed interval $J\subset(0,\beta_0)$,*

$$\limsup_{T\to\infty}\ \sup_{\beta\in J} T^{2\beta/(1+2\beta)} S_T\left(\widehat{X}^{(\beta)}_T;\psi,\beta,R,\delta,\rho,\sigma_+\right)<\infty\,,$$

where S_T is defined by (4.3.6).

The following lemma shows that, given a locally bounded set of minimax-rate predictors, we can always pick a finite subset of at most $N=\lceil(\log T)^2\rceil$ predictors among which the best one achieves the minimax rate of any unknown smoothness index.

Lemma 6. *Let ψ be a distribution on $\mathbb{R}$. Let $\beta_0\in(0,\infty]$ and $\{\widehat{X}^{(\beta)},\beta\in(0,\beta_0)\}$ be a corresponding locally bounded set of ψ-minimax-rate predictors. If $\beta_0<\infty$, select a*

number of predictors $N \geq \lceil \log T \rceil$*, and, in the case where* $\beta_0 = \infty$*, let* $N \geq \lceil (\log T)^2 \rceil$*. Define*

$$\beta_i = \begin{cases} (i-1)\beta_0/N & \text{if } \beta_0 < \infty, \\ (i-1)/N^{1/2} & \text{otherwise,} \end{cases} \qquad 1 \leq i \leq N \,. \tag{4.3.10}$$

Then, we have, for all $\beta \in (0, \beta_0)$*,* $\delta \in (0,1)$*,* $R > 0$*,* $\rho > 0$ *and* $\sigma_+ > 0$*,*

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} \min_{i=1,\dots,N} S_T\left(\widehat{X}_T^{(\beta_i)}; \psi, \beta, R, \delta, \rho, \sigma_+\right) < \infty \,.$$

The proof of this lemma is postponed to Section 4.8.3 in Section 4.8. Lemma 6 says that to obtain a minimax-rate predictor which adapts to an unknown smoothness index β, it is sufficient to select it judiciously among $\log T$ or $(\log T)^2$ well chosen non-adaptive minimax-rate predictors.

As a consequence of Theorem 4.2.1 and Lemma 6, we obtain an adaptive predictor by aggregating them (instead of selecting one of them), as stated in the following result.

Theorem 4.3.2. *Let* ψ *be a distribution on* $\mathbb{R}$*. Let* $\beta_0 \in (0, \infty]$ *and* $\{\widehat{X}^{(\beta)}, \beta \in (0, \beta_0)\}$ *be a locally bounded set of* ψ*-minimax-rate and L-Lipschitz predictors with L satisfying* (L-1)*. Define* $(\widehat{X}_{t,T})_{1\leq t\leq T}$ *as the predictor aggregated from* $\{\widehat{X}^{(\beta_i)}, 1 \leq i \leq N\}$ *with N defined by*

$$N = \begin{cases} \lceil \log T \rceil & \text{if } \beta_0 < \infty, \\ \lceil (\log T)^2 \rceil & \text{otherwise,} \end{cases} \tag{4.3.11}$$

β_i *defined by (4.3.10), and with weights defined according to one of the following setting depending on the assumption on* ψ *and* β_0 *:*

(i) If ψ *satisfies* (I-1) *with* $p \geq 4$ *and* $\beta_0 \leq 1/2$*, use the weights* (4.2.5) *with* $\eta = \sigma_+^{-2}(\log(\lceil \log T \rceil)/T)^{1/2}$*,*

(ii) If ψ *satisfies* (I-1) *with* $p > 2$ *and* $\beta_0 \leq (p-2)/4$*, use the weights* (4.2.6) *with* $\eta = \sigma_+^{-2}(\log(\lceil \log T \rceil)/T)^{2/p}$*,*

(iii) If ψ *satisfies* (I-2)*, use the weights* (4.2.6) *with* $\eta = \sigma_+^{-2}(\log T)^{-3}$*.*

Then, we have, for any $\beta \in (0, \beta_0)$*,* $\delta \in (0,1)$*,* $R > 0$*,* $\rho \in (0,1]$ *and* $\sigma_+ > 0$*,*

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} S_T\left(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+\right) < \infty \,. \tag{4.3.12}$$

The proof of this theorem is postponed to Section 4.8.4 in Section 4.8.

Remark 6. The limitation to $\beta_0 \leq 1/2$ in (i) under Assumption (I-1) for ψ follows from the factor $(\log N/T)^{1/2}$ obtained in the oracle inequality (4.2.8) of Theorem 4.2.1 after optimizing in η (see (4.2.15)). If $p > 4$ this restriction is weakened to $\beta_0 \leq (p-2)/4$ in (ii) taking into account the factor $(\log N/T)^{1-2/p}$ obtained in the oracle inequality (4.2.9) of Theorem 4.2.1 after optimizing in η (see (4.2.17)). In the last case, the limitation of β_0 drops when applying the oracle inequality (4.2.11) of the same theorem. However a stronger condition on ψ is then required.

Remark 7. It may happen that the locally bounded set of ψ-minimax-rate predictors is limited to some $\beta_0 < \infty$ (see the example of the NLMS predictors in Section 4.7.2). In this case, the result roughly needs $\log T$ predictors and the computation of the aggregated one requires less operations than if β_0 were infinite. For these reasons, we do not consider in general that $\beta_0 = \infty$. On the one hand, a finite β_0 yields a restriction on the set of (unknown) smoothness indices β for which the aggregated predictors are minimax rate adaptive. On the other hand, if $\beta_0 = \infty$, Theorem 4.3.2 then requires the stronger Assumption (I-2) on the process.

Remark 8. The constant σ_+^{-2} present in the definitions of η in the three Cases (i), (ii) and (iii) corresponds to the homogenization of the remaining terms appearing in Theorem 4.2.1 (the second lines of (4.2.8), (4.2.9) and (4.2.11)). Indeed with the proposed choices and in the three cases, the constant σ_+^2 factors out in front of the remaining terms (see the last three displayed equations in Section 4.8.4 of Section 4.8). However, the σ_+^{-2} in the definitions of η does not impact the convergence rate in the sense that Theorem 4.3.2 is still valid using any other constant (1 for example) in these definitions.

4.4 PROOFS OF THE UPPER BOUNDS

4.4.1 Proof of Lemma 5

With the weights defined by (2.4.4), we slightly adapt the proof of (Stoltz, 2011, Theorem 1.7). We have that

$$\sum_{t=1}^{T} (\widehat{x}_t - x_t)^2 - \inf_{\nu \in \mathcal{S}_N} \frac{1}{T} \sum_{t=1}^{T} \left(\widehat{x}_t^{[\nu]} - x_t\right)^2 \le 2 \sup_{\nu \in \mathcal{S}_N} \sum_{t=1}^{T} (\widehat{x}_t - x_t)\left(\widehat{x}_t - \widehat{x}_t^{[\nu]}\right)$$
$$\le \sum_{t=1}^{T} \sum_{i=1}^{N} 2\widehat{\alpha}_{i,t} (\widehat{x}_t - x_t) \widehat{x}_t^{(i)} - \min_{i=1,\dots,N} \sum_{t=1}^{T} 2 (\widehat{x}_t - x_t) \widehat{x}_t^{(i)} \,. \quad (4.4.1)$$

For going from the left-hand side to the right-hand side of the first line of Equation (4.4.1) we used the convexity of $x \mapsto x^2$ and for going from the first line to the second one we just lower bounded each term of the sum represented by $\widehat{x}_t^{[\nu]}$. The remainder of the proof relies on Hoeffding's lemma, which we recall. Let X be a random variable with values in $[-B, B]$; then, for all $s \in \mathbb{R}$ we have

$$\log \mathbb{E}\left[\exp\left(sX\right)\right] \le s\mathbb{E}\left[X\right] + \frac{s^2}{2} B^2 \,. \quad (4.4.2)$$

Let $s_t = 2 \max_{1 \le i \le N} |2(\widehat{x}_t - x_t)\widehat{x}_t^{(i)}|$. For each $t = 1, \dots, T$, we apply (4.4.2) to a random variable taking the value $2(\widehat{x}_t - x_t)\widehat{x}_t^{(i)}$ with probability $\widehat{\alpha}_{i,t}$. Observe that its absolute value is upper bounded by $s_t/2$. Setting $s = -\eta$ and summing all these inequalities yield

$$\eta \sum_{t=1}^{T} \sum_{i=1}^{N} 2\widehat{\alpha}_{i,t} (\widehat{x}_t - x_t) \widehat{x}_t^{(i)} \leq -\log \prod_{t=1}^{T} \frac{\sum_{i=1}^{N} \widehat{\alpha}_{i,t} \exp\left(-2\eta \sum_{s=1}^{t} (\widehat{x}_s - x_s) \widehat{x}_s^{(i)}\right)}{\sum_{i=1}^{N} \widehat{\alpha}_{i,t} \exp\left(-2\eta \sum_{s=1}^{t-1} (\widehat{x}_s - x_s) \widehat{x}_s^{(i)}\right)} + \frac{\eta^2}{8} s_T^*$$
$$= -\log \sum_{i=1}^{N} \widehat{\alpha}_{i,t} \exp\left(-2\eta \sum_{s=1}^{T} (\widehat{x}_s - x_s) \widehat{x}_s^{(i)}\right) + \frac{\eta^2}{8} s_T^* ,$$

where $s_T^* = \sum_{t=1}^{T} s_t^2$. The bound (4.2.12) follows by lower bounding the sum of the exponential terms in the logarithm of the right-hand side by the largest of them and by observing that $s_t^2 \leq 16 y_t^2$.

We now prove (4.2.13). We adapt the proof of (Catoni, 2004, Proposition 2.2.1) to unbounded sequences by replacing the convexity argument by the following lemma.

Lemma 7. *Let $a > 0$ and $\mathbb{P}$ a probability distribution supported on $[-a, a]$. Then we have*

$$\int \exp\left(-x^2\right) \mathrm{d}\mathbb{P}(x) \leq \exp\left(-\left(\int x \mathrm{d}\mathbb{P}(x)\right)^2 + \left(a^2 - \frac{1}{2}\right)_+\right) .$$

The proof of Lemma 7 is postponed to Section 4.8.5 in Section 4.8. Now, let $\eta > 0$ and $t = 1, \ldots, T$. Using Lemma 7 with the probability distribution $\mathbb{P}$ defined by $\mathbb{P}(A) = \sum_{i=1}^{N} \widehat{\alpha}_{i,t} \mathbb{1}_A(\eta^{1/2}(\widehat{x}_t^{(i)} - x_t))$ and $a = \eta^{1/2} y_t$, we get that

$$\sum_{i=1}^{N} \widehat{\alpha}_{i,t} \exp\left(-\eta \left(\widehat{x}_t^{(i)} - x_t\right)^2\right) \leq \exp\left(-\eta (\widehat{x}_t - x_t)^2 + \eta \left(y_t^2 - \frac{1}{2\eta}\right)_+\right) .$$

Taking the log, multiplying by $-\eta^{-1}$ and re-ordering the terms, we obtain that

$$(\widehat{x}_t - x_t)^2 \leq -\frac{1}{\eta} \log\left(\sum_{j=1}^{N} \widehat{\alpha}_{i,t} \exp\left(-\eta \left(\widehat{x}_t^{(i)} - x_t\right)^2\right)\right) + \left(y_t^2 - \frac{1}{2\eta}\right)_+ .$$

Taking the average over $t = 1, \ldots, T$ and developing the expression of $\widehat{\alpha}_{i,t}$, we obtain

$$\frac{1}{T} \sum_{t=1}^{T} (x_t - \widehat{x}_t)^2 \leq -\frac{1}{\eta T} \log\left(\frac{1}{N} \sum_{i=1}^{N} \exp\left(-\eta \sum_{t=1}^{T} \left(\widehat{x}_t^{(i)} - x_t\right)^2\right)\right)$$
$$+ \frac{1}{T} \sum_{t=1}^{T} \left(y_t^2 - \frac{1}{2\eta}\right)_+ . \quad (4.4.3)$$

Using that $\sum_{i=1}^{N} \exp(-\eta \sum_{t=1}^{T} (\widehat{x}_t^{(i)} - x_t)^2) \geq \exp(-\eta \min_{i=1,\ldots,N} \sum_{t=1}^{T} (\widehat{x}_t^{(i)} - x_t)^2)$, we get the bound (4.2.13).

4.4.2 Proof of Theorem 4.2.1

We prove the Cases (i), (ii) and (iii) successively. We denote $Y_t = |X_t| + \max_{1\le i\le N} |\widehat{X}_t^{(i)}|$.
Case (i). Applying (4.2.12) in Lemma 5 with $\mathbb{E}[\inf \dots] \le \inf \mathbb{E}[\dots]$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{[\nu]} - X_t\right)^2\right] + \frac{\log N}{T\eta} + \frac{2\eta}{T}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^4\right] . \quad (4.4.4)$$

Using that the predictors are L-Lipschitz and the process $(X_t)_{t\in\mathbb{Z}}$ satisfies (M-1), we have, for all $1 \le t \le T$,

$$Y_t = |X_t| + \max_{1\le i\le N}\left|\widehat{X}_t^{(i)}\right| \le \sum_{j\in\mathbb{Z}} A_t(j)\, Z_{t-j} + \sum_{s\ge 1}\sum_{j\in\mathbb{Z}} L_s\, A_{t-s}(j)\, Z_{t-s-j} \le \sum_{j\in\mathbb{Z}} B_t(j) Z_{t-j}, \quad (4.4.5)$$

where

$$B_t(j) = A_t(j) + \sum_{s\ge 1} L_s\, A_{t-s}(j-s) .$$

Applying the Minkowski inequality together with (4.4.5), (4.2.2) and (4.2.7), we obtain, for all $1 \le t \le T$,

$$\mathbb{E}\left[Y_t^4\right] \le \mathbb{E}\left[\left(\sum_{j\in\mathbb{Z}} B_t(j) Z_{t-j}\right)^4\right] \le A_*^4 (1+L_*)^4 \sup_{t\in\mathbb{Z}}\mathbb{E}\left[Z_t^4\right] .$$

Since the process Z fulfills (N-1) with $p = 4$, plugging this bound in (4.4.4) we obtain (4.2.8).
Case (ii). We use (4.2.13) in Lemma 5 and the inequality $(x^2 - 1/(2\eta))_+ \le (2\eta)^{p/2-1} x^p$ which holds for $x \ge 0$ and $p \ge 2$. We get, taking the expectation,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] \le \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T}^{(i)} - X_{t,T}\right)^2\right] + \frac{\log N}{T\eta} + (2\eta)^{p/2-1}\max_{t=1,\dots,T}\mathbb{E}\left[Y_t^p\right] . \quad (4.4.6)$$

Applying the Minkowski inequality, (4.4.5) and Assumption (N-2)

$$\mathbb{E}\left[Y_t^p\right] \le \left(\sum_{j\in Z} B_t(j)\left(\mathbb{E}\left[Z_{t-j}^p\right]\right)^{1/p}\right)^p \le A_*^p (1+L_*)^p \sup_{t\in Z}\mathbb{E}\left[Z_t^p\right] .$$

Using this bound which is independent of t, with (N-1) and (4.4.6), the inequality (4.2.9) follows.
Case (iii). To obtain (4.2.11), we again use (4.2.13) in Lemma 5 but now with an exponential bound for $(Y_t^2 - 1/(2\eta))_+$. We note that, or all $u > 0$,

$$\sup_{x\ge 1}\left(x^2-1\right)\mathrm{e}^{-ux} = \left(x_0^2-1\right)\mathrm{e}^{-u\,x_0} \quad \text{with} \quad x_0 = u^{-1}\left(1+\left(1+u^2\right)^{1/2}\right) .$$

It follows that, for all $x \in \mathbb{R}$ and $u > 0$,

$$\left(x^2 - 1\right)_+ \le \mathrm{e}^{ux}\left(x_0^2 - 1\right)\mathrm{e}^{-u\,x_0} \le \mathrm{e}^{ux} 2u^{-2}\,(2+u)\,\mathrm{e}^{-1-u}\,.$$

Applying this bound with $x = (2\eta)^{1/2} Y_t$ and $u = \lambda(2\eta)^{-1/2}$ we get

$$\left(Y_t^2 - \frac{1}{2\eta}\right)_+ = (2\eta)^{-1}\left(x^2-1\right)_+ \le 2\lambda^{-2}\left(2 + \lambda(2\eta)^{-1/2}\right)\mathrm{e}^{-1-\lambda(2\eta)^{-1/2}}\,\mathrm{e}^{\lambda Y_t}\,.$$

Plugging this into (4.2.13) and taking the expectation, we obtain that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] \le \min_{i=1,\dots,N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}^{(i)}_{t,T} - X_{t,T}\right)^2\right] + \frac{\log N}{T\eta} \\ + 2\lambda^{-2}\left(2+\lambda(2\eta)^{-1/2}\right)\mathrm{e}^{-1-\lambda(2\eta)^{-1/2}}\max_{t=1,\dots,T}\mathbb{E}\left[\mathrm{e}^{\lambda Y_t}\right]\,. \quad (4.4.7)\end{aligned}$$

We now use Assumption (**N**-2). Since $B_t(j) \le a^*(1+L_*)$ for all $j, t \in \mathbb{Z}$ and

$$\sum_{j\in\mathbb{Z}} B_t(j) \le A_*(1+L_*),$$

Jensen's inequality and (4.4.5) gives that, for any $\lambda \le \zeta/(a^*(1+L_*))$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{\lambda Y_t}\right] &\le \mathbb{E}\left[\mathrm{e}^{\lambda\left(|X_t| + \max_{1\le i\le N}|\widehat{X}^{(i)}_t|\right)}\right] \\ &\le \prod_{j\in\mathbb{Z}}\mathbb{E}\left[\mathrm{e}^{\lambda B_t(j) Z_{t-j}}\right] \\ &\le \prod_{j\in\mathbb{Z}}(\phi(\zeta))^{\lambda B_t(j)/\zeta} \le (\phi(\zeta))^{\lambda A_*(1+L_*)/\zeta}\,.\end{aligned}$$

The combination of this bound with (4.4.7) gives (4.2.11). The proof of Theorem 4.2.1 is complete.

4.4.3 Proof of Case (iii) in Corollary 1

Minimizing the sum of the two terms appearing in the second line of (4.2.11) is a bit more involved, since it depends both on η and λ. Under Condition (4.2.10), the quantity $(\phi(\zeta))^{\lambda A_*(1+L_*)/\zeta}$ remains between two positive constants while, for any $\eta > 0$, $\lambda^{-2}(2 + \lambda(2\eta)^{-1/2})$ is decreasing as λ increases. To simplify $(\phi(\zeta))^{\lambda A_*(1+L_*)/\zeta}$ into $\phi(\zeta)$, we simply take

$$\lambda = \frac{\zeta}{A_*(1+L_*)}\,,$$

which satisfies (4.2.10). Now that λ is set, it remains to choose a value of η which (almost) minimizes

$$\frac{\log N}{T\eta} + \frac{2\phi(\zeta)}{\mathrm{e}}\lambda^{-2}\left(2+\lambda(2\eta)^{-1/2}\right)\mathrm{e}^{-\lambda(2\eta)^{-1/2}}\,.$$

The η defined as in (4.2.18) is chosen so that $(\log N)/T = \mathrm{e}^{-\lambda(2\eta)^{-1/2}}$, and we get (4.2.19).

4.5 PROOF OF THE LOWER BOUND

We now provide a proof of Theorem 4.3.1. We consider an autoregressive equation of order one

$$X_{t,T} = \theta\left(\frac{t-1}{T}\right) X_{t-1,T} + \xi_t, \tag{4.5.1}$$

where $(\xi_t)_{t\in\mathbb{Z}}$ is i.i.d. with density f as in (I-3). In this case, if $\sup_{u\le 1} |\theta(u)| < 1$, the representation (4.3.4) of the stationary solution reads, for all $t \le T$ as

$$X_{t,T} = \sum_{j=0}^{\infty} \prod_{s=1}^{j} \theta\left(\frac{t-s}{T}\right) \xi_{t-j}\,, \tag{4.5.2}$$

with the convention $\prod_{s=1}^{0} \theta((t-s)/T) = 1$. The class of models so defined with $\theta \in \Lambda_1(\beta, R) \cap s_1(\delta)$ corresponds to Assumption (M-2) with $(\boldsymbol{\theta}, \sigma)$ in $C(\beta, R, \delta, \rho, 1)$ such that only the first component of $\boldsymbol{\theta}$ is nonzero and σ is constant and equal to one.

We write henceforth in this proof section $\mathbb{P}_\theta$ for the law of the process $X = (X_{t,T})_{t\le T, T\ge 1}$ and $\mathbb{E}_\theta$ for the corresponding expectation.

Let $\widehat{X} = (\widehat{X}_{t,T})_{1\le t\le T}$ be any predictor of $(X_{t,T})_{1\le t\le T}$ in the sense of Definition 9. Define $\widehat{\theta} = (\widehat{\theta}_{t,T})_{0\le t\le T-1} \in \mathbb{R}^T$ by

$$\widehat{\theta}_{t,T} = \begin{cases} \widehat{X}_{t+1,T}/X_{t,T} & \text{if } X_{t,T} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

For any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^T$, we define

$$d_X(\mathbf{u}, \mathbf{v}) = \left(\frac{1}{T}\sum_{t=0}^{T-1} X_{t,T}^2\,(u_t - v_t)^2\right)^{1/2}. \tag{4.5.3}$$

By (4.5.1), since $X_{t,T}$ and $\widehat{\theta}_{t,T}$ are $\mathcal{F}_{t,T}$-measurable, they are independent of ξ_{t+1} and we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_\theta\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] - 1 = \mathbb{E}_\theta\left[d_X^2(\widehat{\theta}, \mathrm{v}_T\{\theta\})\right],$$

where, for any $\theta : (-\infty, 1] \to \mathbb{R}$, $\mathrm{v}_T\{\theta\} \in \mathbb{R}^T$ denotes the T-sample of θ on the regular grid $0, 1/T, \ldots, (T-1)/T$,

$$\mathrm{v}_T\{\theta\} = \left(\theta\left(\frac{t}{T}\right)\right)_{0\le t\le T-1}.$$

Hence to prove the lower bound of Theorem 4.3.1, it is sufficient to show that there exist $\theta_0, \ldots, \theta_M \in \Lambda_1(\beta, R) \cap s_1(\delta)$, $c > 0$ and $T_0 \ge 1$ both depending only on δ, β, R and the density f, such that for any $\widehat{\theta} = (\widehat{\theta}_{t,T})_{0\le t\le T-1}$ adapted to $(\mathcal{F}_{t,T})_{0\le t\le T-1}$ and $T \ge T_0$, we have

$$\max_{j=0,\ldots,M} \mathbb{E}_{\theta_j}\left[d_X^2\left(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}\right)\right] \ge c\, T^{-2\beta/(2\beta+1)}. \tag{4.5.4}$$

We now face the more standard problem of providing a lower bound for the minimax rate of an estimation error, since $\widehat{\theta}$ is an estimator of $\mathrm{v}_T\{\theta\}$. The path for deriving such a lower bound is explained in (Tsybakov, 2009, Chapter 2). However we have to deal with a loss function d_X which depends on the observed process X. Not only the loss function is random, but it is also not independent of the estimator $\widehat{\theta}$. The proof of the lower bound (4.5.4) thus requires non-trivial adaptations. It relies on some intermediate lemmas.

Lemma 8. *We write $\mathcal{K}(\mathbb{P}, \mathbb{P}')$ for the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{P}'$. For any functions $\theta_0, \dots, \theta_M$ from $[0,1]$ to $\mathbb{R}$ such that*

$$\max_{j=0,\dots,M} \mathcal{K}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}) \leq \frac{2\mathrm{e}}{2\mathrm{e}+1} \log(1+M) \tag{4.5.5}$$

and any $r > 0$ we have

$$\max_{j=0,\dots,M} \mathbb{E}_{\theta_j}\left[d_X^2\left(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}\right)\right] \geq \frac{r^2}{4}\left(\frac{1}{2\mathrm{e}+1} - \max_{j=0,\dots,M} \mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} d_{X,T}(\theta_i, \theta_j) \leq r\right)\right),$$

where we denote, for any two functions θ, θ' from $(-\infty, 1]$ to $\mathbb{R}$,

$$d_{X,T}(\theta, \theta') = d_X\left(\mathrm{v}_T\{\theta\}, \mathrm{v}_T\{\theta'\}\right) .$$

The proof is postponed to Section 4.8.6 in Section 4.8.
We next construct certain functions $\theta_0, \dots, \theta_M \in \Lambda_1(\beta, R) \cap s_1(\delta)$ fulfilling (4.5.5) and well spread in terms of the pseudo-distance $d_{X,T}$. Consider the infinitely differentiable kernel K defined by

$$K(u) = \exp\left(-\frac{1}{1-4u^2}\right) \mathbb{1}_{|u|<1/2} .$$

Given any $m \geq 8$, Vershamov-Gilbert's lemma (Tsybakov, 2009, Lemma 2.9) ensures the existence of $M+1$ points $w^{(0)}, \dots, w^{(M)}$ in the hypercube $\{0,1\}^m$ such that

$$M \geq 2^{m/8}, \ w^{(0)} = 0 \quad \text{and} \quad \mathrm{card}\left\{\ell : w_\ell^{(j)} \neq w_\ell^{(i)}\right\} \geq m/8 \quad \text{for all } j \neq i. \tag{4.5.6}$$

We then define $\theta_0, \dots, \theta_M$ by setting, for all $x \leq 1$,

$$\theta_j(x) = \frac{R_0}{m^\beta} \sum_{\ell=1}^{m} w_l^{(j)} K\left(mx - \ell + \frac{1}{2}\right) \quad \text{for} \quad j = 0, \dots, M , \tag{4.5.7}$$

where

$$R_0 = \min\left(\delta, \frac{R}{\left(2\,|K|_\beta\right)}\right) . \tag{4.5.8}$$

Since $K = 0$ out of $(-1/2, 1/2)$, we observe that

$$\theta_j(x) = 0 , \quad \text{for all } x \leq 0 , \tag{4.5.9}$$

and

$$\theta_j(x) = \frac{R_0}{m^\beta}\, w^{(j)}_{\lfloor mx\rfloor+1}\, K\left(\{mx\} - \frac{1}{2}\right), \quad \text{for all } x \in [0,1], \tag{4.5.10}$$

where $\{mx\} = mx - \lfloor mx\rfloor$ denotes the fractional part of mx. Thus, we have

$$\theta^* := \max_{0\le j\le M} \sup_{x\in[0,1]} |\theta_j(x)| \le \frac{R_0 \mathrm{e}^{-1}}{m^\beta} \le \delta < 1\ . \tag{4.5.11}$$

We first check that the definition of R_0 ensures that the θ_j's are in the expected set of parameters.

Lemma 9. *For all $j = 0, \dots, M$, we have $\theta_j \in \Lambda_1(\beta, R) \cap s_1(\delta)$.*

The proof can be found in Section 4.8.7 of Section 4.8.
Next we provide a bound to check the required condition (4.5.5) on the chosen θ_j's.

Lemma 10. *For all $j = 1, \dots, M$, we have*

$$\mathcal{K}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}) \le \frac{8\,\mathrm{e}^{-2}\,\kappa\, R_0^2}{(1-\delta^2)\log 2}\ \frac{T}{m^{1+2\beta}}\log(1+M)\ ,$$

where κ is the constant appearing in (I-3).

We prove it in Section 4.8.8 of Section 4.8.
Finally we need a control on the distances $d^2_{X,T}(\theta_i, \theta_j)$.

Lemma 11. *For any $\varepsilon > 0$, there exists a constant A depending only on ε and the density f of ξ such that for all $m \ge 16$, $T \ge 4m$ and $j = 0, \dots, M$,*

$$\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} d^2_{X,T}(\theta_i, \theta_j) \le A\frac{R_0^2}{m^{2\beta}}\right) \le \varepsilon + \frac{2R_0\mathrm{e}^{-3}}{A\,(1-\delta)m^\beta}\ . \tag{4.5.12}$$

The proof is postponed to Section 4.8.9 of Section 4.8.
We can now complete the proof of Theorem 4.3.1.

Proof of Theorem 4.3.1. Recall that $\theta_0, \dots, \theta_M$ in (4.5.7) are some parameters only depending on β and δ and a certain integer $m \ge 8$ and that, whatever the value of m, Lemma 9 insures that $\theta_0, \dots, \theta_M$ belongs to $\Lambda_1(\beta, R) \cap s_1(\delta)$.
Hence it is now sufficient to show that (4.5.4) holds for a correct choice of m, relying on Lemmas 8, 10 and 11. Let us set

$$m = \max\left\{\left\lceil c_0 T^{1/(2\beta+1)}\right\rceil, 16\right\}\ , \tag{4.5.13}$$

where c_0 is a constant to be chosen. Then $Tm^{-1-2\beta} \le c_0^{-1-2\beta}$ and, by Lemma 10, we can choose c_0 only depending on β, R, κ and δ so that Condition (4.5.5) of Lemma 8 is met. We thus get that, for any $r > 0$,

$$\max_{j=0,\dots,M} \mathbb{E}_{\theta_j}\left[d^2_X\left(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}\right)\right] \ge \frac{r^2}{4}\left(\frac{1}{2\mathrm{e}+1} - \max_{j=0,\dots,M}\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} d_{X,T}(\theta_i,\theta_j) \le r\right)\right),$$

Applying Lemma 11 with $\varepsilon = 1/(4\mathrm{e}+2)$ and the previous bound with $r^2 = A R_0^2 m^{-2\beta}$, we get, as soon as $T \geq 4m$,

$$\max_{j=0,\dots,M} \mathbb{E}_{\theta_j}\left[d_X^2\left(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}\right)\right] \geq \frac{r^2}{4}\left(\frac{1}{4\mathrm{e}+2} - \frac{2R_0\mathrm{e}^{-1}}{A\,(1-\delta)m^{\beta}}\right) .$$

The proof is concluded by observing that, as a consequence of (4.5.13), we can choose a constant T_0 only depending on β, R, κ and δ such that $T \geq T_0$ implies that $T \geq 4m$ and that the term between parentheses is bounded by $1/(8\mathrm{e}+4)$ from below. □

4.6 NUMERICAL EXPERIMENTS

In this section, we test the proposed aggregation methods on data simulated according to a TVAR process with $d = 3$. The choice of a smooth parameter function $t \mapsto \boldsymbol{\theta}(t)$ within $s_d(\delta)$ for some $\delta \in (0,1)$ is done by first picking randomly some smoothly time varying partial autocorrelation functions up to the order d that are bounded between -1 and 1 and then by relying on the Levinson-Durbin algorithm. We show the three components of the obtained $\boldsymbol{\theta}(t)$ on $t \in [0,1]$ in the top parts of Figure 4.1. Realizations of the TVAR process are then obtained from an innovation sequence $(\xi_t)_{t\in\mathbb{Z}}$ of i.i.d. centered Gaussian process with unit variance as in Definition 8 by sampling $\boldsymbol{\theta}$ at a given rate $T \geq 1$. Figure 4.1 displays one realization of such a TVAR process for $T = 2^{10}$.

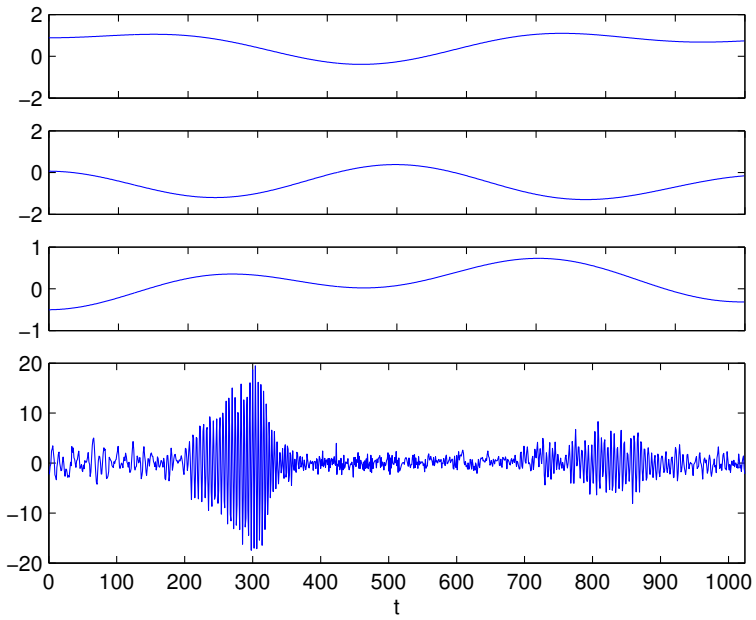


Figure 4.1 : The first three plots represent θ_1, θ_2 and θ_3 on the interval $[0,1]$. The last plot displays $T = 2^{10}$ samples of the corresponding TVAR process with Gaussian innovations.

The NLMS algorithm (see Algorithm 6 in Section 4.7.2) studied in Moulines et al. (2005) provides an online estimator of $\boldsymbol{\theta}$ depending on a gradient step size μ. For any $\beta \in (0,1]$

and any constant $c > 0$, choosing $\mu = c\,T^{-2\beta/(2\beta+1)}$ yields a $C(\beta, R, \delta, \rho, 1)$–minimax-rate online L-Lipschitz predictor as explained in Section 4.7. Hence, taking $c = 0.01$ (we selected its value within the set $\{1, 0.1, 0.05, 0.01, 0.005, 0.001\}$ to better illustrate our theoretical results) and proceeding as in Lemma 6 to define N and β_i, $i = 1, \ldots, N$, with $\beta_0 = 0.5$, we obtain a finite set of NLMS predictors corresponding to gradient step sizes $\mu_1 > \cdots > \mu_N$. This set of predictors is aggregated in two possible ways according to the online Algorithm 5 with the specifications on η and N given in Theorem 4.3.2. The overall running time of T iterates of the algorithm leading to the aggregated predictors from the data $X_1, \ldots, X_T$ is then $O(d\,N\,T)$. Since the algorithm is recursive, the corresponding required storage capacity is $O(d\,N)$.
We evaluate the obtained NLMS predictors and their aggregated predictors by running 1000 simulations based on equally distributed realizations of the above Gaussian TVAR process in the case $T = 2^{10}$ which yields $N = 7$. In Figure 4.2 we compare the averaged downward shifted empirical losses defined for any predictor $(\widehat{X}_{t,T})_{1\le t\le T}$ by

$$L_T = \frac{1}{T}\sum_{t=1}^{T}\left(\left(\widehat{X}_{t,T} - X_{t,T}\right)^2 - \sigma^2\left(\frac{t}{T}\right)\right).$$

This empirical averaged loss mimics the risk considered in (4.3.6).

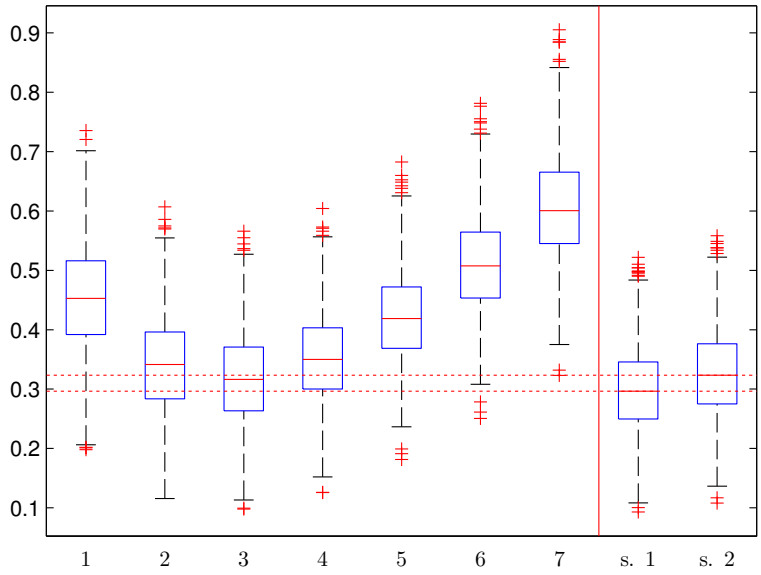


Figure 4.2 : The seven boxplots on the left of the vertical red line correspond to the averaged downward shifted empirical losses L_T of the NLMS predictors $\widehat{X}^{(1)}, \ldots, \widehat{X}^{(7)}$. The ones on the right of the same line are those associated with the aggregated predictors using the weights (4.2.5) and (4.2.6).

We observe that the best NLMS predictor is the third one while the aggregated predictor of Strategy 1 enjoys a smaller loss and that of Strategy 2 a slightly larger one. This is in accordance with Theorem 4.2.1 (i) and (iii) where it is shown that the aggregated predictor

of the first strategy may outperform the best predictor as it nearly achieves the loss of the best possible convex combination of the original predictors while the aggregated predictor of the second strategy nearly achieves the loss of the best original predictor.

4.7 Application to online minimax adaptive prediction

4.7.1 From estimation to prediction

We define a sequence $(L_k)_{k\geq 1}$ by

$$L_k = \begin{cases} \binom{d}{k} & \text{if } 1 \leq k \leq d \\ 0 & \text{otherwise,} \end{cases}$$

which fulfills (L-1) with $L_* = \sum_{k=1}^{d} \binom{d}{k} = 2^d - 1$. Given an estimator $\widehat{\boldsymbol{\theta}}_{t-1,T} = [\widehat{\theta}_{t-1,T}(1) \dots \widehat{\theta}_{t-1,T}(d)]'$, we define a predictor $\widehat{X}_{t,T}$ which is L-Lipschitz by setting

$$\widehat{X}_{t,T} = \sum_{k=1}^{d} \left(\min \left\{ \max \left\{ -L_k, \widehat{\theta}_{t-1,T}(k) \right\}, L_k \right\} \right) X_{t-k,T} \,. \tag{4.7.1}$$

The predictor $\widehat{X}_{t,T}$ is the natural linear predictor $\widehat{\boldsymbol{\theta}}'_{t-1,T} \mathbf{X}_{t-1,T}$, where A' denotes the transpose of matrix A and $\boldsymbol{X}_{s,T} = [X_{s,T} \dots X_{s-(d-1),T}]'$, normalized to be at most L-Lipschitz. The normalization step amounts to project $\widehat{\boldsymbol{\theta}}_{t,T}$ on a rectangle $[-L_1, L_1] \times \dots \times [-L_d, L_d]$ before deriving the linear predictor. This can only improve the quality of estimation for a stable TVAR model, since $\boldsymbol{\theta}$ takes values in the maximal set of stability $s_d(1)$, which implies that it is included in this rectangle at every point, see (Moulines et al., 2005, Equation 12). We get the following result.

Lemma 12. *Suppose that Assumption (M-2) holds. Consider, for some $1 \leq t \leq T$, an estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_{t,T})_{0\leq t\leq T-1}$ adapted to the filtration $(\mathcal{F}_{t,T})_{0\leq t\leq T-1}$. Define a predictor $\widehat{X} = (\widehat{X}_{t,T})_{1\leq t\leq T}$ as in (4.7.1). Then, for any $q > 1$ and for all and $1 \leq t \leq T$,*

$$\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)} \left[\left(\widehat{X}_{t,T} - X_{t,T} \right)^2 \right] - \sigma^2 \left(\frac{t}{T} \right) \leq C_T(q) \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)} \left[\left| \widehat{\boldsymbol{\theta}}_{t-1,T} - \boldsymbol{\theta}_{t-1,T} \right|^{2q} \right] \right)^{1/q} , \tag{4.7.2}$$

where

$$C_T(q) = \max_{1\leq t\leq T} \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)} \left[\left| \mathbf{X}_{t-1,T} \right|^{2q'} \right] \right)^{1/q'} ,$$

with $1/q' + 1/q = 1$.

Remark 9. Assume that the distribution ψ of the innovations satisfies (I-1) for some $p \geq 2q' > 2$. Then, the Proposition 2 combined with the Minkowski inequality ensure that there exists T_0, $\bar{K}$, δ_1 such that, for any $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, 0, \sigma_+)$,

$$C_T(q) \leq d \left(\frac{\bar{K}\sigma_+}{1-\delta_1} \right)^2 m_{2q'}^{1/q'} , \quad \text{for all } T \geq T_0 \,.$$

Proof. Denote by $\widetilde{\boldsymbol{\theta}}_{t,T}$ the projection of $\widehat{\boldsymbol{\theta}}_{t,T}$ onto the rectangle $[-L_1, L_1] \times \cdots \times [-L_d, L_d]$, that is, $\widetilde{\theta}_{t,T}(k) = \min\{\max\{-L_k, \widehat{\theta}_{t,T}(k)\}, L_k\}$. By (Moulines et al., 2005, Equation 12), $\boldsymbol{\theta}_{t,T}$ lies in this rectangle and thus

$$\left|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}_{t,T}\right| \leq \left|\widehat{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}_{t,T}\right| . \tag{4.7.3}$$

Using (4.8.5) and that $\widehat{\boldsymbol{\theta}}_{t-1,T}$ is a $\mathcal{F}_{t-1,T}$-measurable, we have, for all $t = 1, \ldots, T$,

$$\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left(\widehat{X}_{t,T} - X_{t,T}\right)^2\right] = \mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)}\left[\left(\left(\widetilde{\boldsymbol{\theta}}_{t-1,T} - \boldsymbol{\theta}_{t-1,T}\right)' \mathbf{X}_{t-1,T}\right)^2\right] + \sigma^2\left(\frac{t}{T}\right) .$$

Define q' by the relation $1/q' + 1/q = 1$. Thus, with (4.7.3) and the Hölder inequality, we get that the left-hand side of (4.7.2) is bounded from above by

$$\left(\mathbb{E}_{(\boldsymbol{\theta},\sigma)}\left[\left|\widehat{\boldsymbol{\theta}}_{t-1,T} - \boldsymbol{\theta}_{t-1,T}\right|^{2q}\right]\right)^{1/q} \left(\mathbb{E}_{(\boldsymbol{\theta},\sigma)}\left[\left|\mathbf{X}_{t-1,T}\right|^{2q'}\right]\right)^{1/q'}$$

which concludes the proof of Lemma 12. □

By Lemma 12, to exhibit (ψ,β)-minimax-rate predictors in the sense of Definition 10, it suffices to have (ψ,β)-minimax-rate estimators of $\boldsymbol{\theta}$ in the sense of L^q-norm. We recall some known results in this direction in the following section, with a focus on online procedures.

4.7.2 Online estimators

Parameter estimation for TVAR models, or, more generally for locally stationary processes has been intensively studied in the past two decades, see Dahlhaus (2009) for a recent overview on this problem. To our knowledge, minimax-rate estimation results are sparse. The more widely spread approach for studying the behaviour of such estimators consists in establishing a central limit theorem under differentiablity conditions. Moment upper bounds are provided in Dahlhaus and Giraitis (1998) and could be used to obtain minimax rate results. However the estimator, which is based on a localized Yule-Walker estimation method is not naturally adapted to the filtration $(\mathcal{F}_{t,T})_{0\leq t\leq T-1}$ as required for $(\widetilde{\boldsymbol{\theta}}_{t,T})_{0\leq t\leq T-1}$ above. Such a constraint could clearly be met with some adaptation of the Yule-Walker approach. On the other hand it is directly satisfied by the estimators studied in Moulines et al. (2005). There, an online estimator is proposed, the normalized least mean squares (NLMS) estimator $\widehat{\boldsymbol{\theta}}_{t,T}(\mu)$, depending on a gradient step size μ. For the sake of completeness, we present the computation of the NLMS estimator in Algorithm 6.

Algorithm 6: Online computation of the NLMS estimator.

parameters the gradient step size μ;
initialization $t = 0$, $\widehat{\boldsymbol{\theta}}_{t,T}(\mu) = \begin{bmatrix} 0 & \dots & 0 \end{bmatrix}'$;
while input *a new* $X_{t,T}$*;*
do

$$\widehat{\boldsymbol{\theta}}_{t,T}(\mu) = \widehat{\boldsymbol{\theta}}_{t-1,T}(\mu) + \mu \left(X_{t,T} - \widehat{\boldsymbol{\theta}}'_{t-1,T}(\mu) \boldsymbol{X}_{t-1,T} \right) \frac{\boldsymbol{X}_{t-1,T}}{1 + \mu \left| \boldsymbol{X}_{t-1,T} \right|^2};$$

return $\widehat{\boldsymbol{\theta}}_{t,T}(\mu)$*;*
$t = t + 1$;

For any $\beta \in (0, 1]$, provided that the gradient step μ is well chosen, the NLMS estimator is (ψ, β)-minimax-rate, see (Moulines et al., 2005, Corollary 3). More precisely, assume (**M-2**) with ψ satisfying (**I-1**) for some $p \geq 4$. Then, for any $c > 0$, $\varepsilon > 0$, $R > 0$, $\delta \in (0, 1)$, $\rho \in [0, 1]$ and $q \in [1, p/6)$, there exists $M > 0$ such that, for all $(\boldsymbol{\theta}, \sigma) \in \mathcal{C}(\beta, R, \delta, \sigma_-, \sigma_+)$ and $\varepsilon > 0$,

$$\sup_{\varepsilon \leq t/T \leq 1} \left(\mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)} \left[\left| \widehat{\boldsymbol{\theta}}_{t,T}(cT^{-2\beta/(1+2\beta)}) - \boldsymbol{\theta}_{t,T} \right|^{2q} \right] \right)^{1/q} \leq M \, T^{-2\beta/(1+2\beta)} .$$

Clearly, from Moulines et al. (2005), the constant M can be bounded uniformly for β in any compact subinterval away from 0, as required in Definition 11. Lemma 12 applies for $q \geq p/(p-2)$ so to meet the condition $q \in [1, p/6)$, we set $q = p/(p-2)$ and impose $p > 8$ and finally obtain that

$$\sup_{\varepsilon \leq t/T \leq 1} \mathbb{E}^{\psi}_{(\boldsymbol{\theta},\sigma)} \left[\left(\widehat{X}_{t,T}(cT^{-2\beta/(1+2\beta)}) - X_{t,T} \right)^2 \right] - \sigma^2 \left(\frac{t}{T} \right) \leq C' \, \sigma_+^2 \, T^{-2\beta/(1+2\beta)} ,$$

where $\widehat{X}_{t,T}(\mu)$ is the predictor defined from the estimator $\widehat{\boldsymbol{\theta}}_{t,T}(\mu)$ as in (4.7.1). This is almost what is required in our Definition 11 except that in (4.3.9) we have $T^{-1} \sum_{t=1}^{T}(\dots)$ instead of $\sup_{\varepsilon \leq t/T \leq 1}(\dots)$. In fact one can take $\varepsilon = 0$, provided that a burn-in period of observation is assumed prior to the time origin. It would only require the NLMS estimator to be running from observations $X_{t,T}$ started at times $t \geq -\varepsilon T$ for some positive ε, which seems a reasonable assumption in practice. Finally, let us recall that, as shown in Moulines et al. (2005), NLMS estimators are no longer minimax rate for a Hölder smoothness index $\beta > 1$. However, a bias reduction technique can be used to obtain a minimax-rate estimator for $\beta \in (1, 2]$, see (Moulines et al., 2005, Corollary 9).
To the best of our knowledge, there are not available minimax-rate estimators for $\beta > 2$. Chapter 5 proposes to fill this gap by relying on an adaptation of the Yule-Walker method.

4.8 Postponed proofs

4.8.1 A useful lemma

The following lemma provides a uniform bound on the norm of a product of matrices sampled from a continuous function defined on an interval I and valued in a set of $d \times d$

matrices with bounded spectral radius and norm.

Lemma 13. *Let $d \geq 1$ and I an interval of $\mathbb{R}$. Let A be a function defined on I taking values in the set of $d \times d$ matrices with eigenvalues moduli at most equal to δ. Let $|\cdot|$ be any matrix norm. Denote by A^* the corresponding uniform norm of A,*

$$A^* = \sup_{t \in I} |A(t)| \ ,$$

and, for any $h > 0$, $\omega_h(A, I)$ the modulus of continuity of A over I,

$$\omega_h(A; I) = \sup \{|A(t) - A(s)| \ : \ s, t \in I, \ |s - t| \leq h\} \ .$$

Let $\delta_1 > \delta$ and assume that $A^* < \infty$. Then there exist some positive constants ε, ℓ and K only depending on A^*, δ and δ_1 such that, for any $h \in (0, 1)$ fulfilling $\omega_h(A; I) \leq \varepsilon$, we have, for all $s < t$ in I and all integer $p \geq \ell(t - s)/h$,

$$\left| \underbrace{A(t) A\left(t - \frac{t-s}{p}\right) A\left(t - \frac{2(t-s)}{p}\right) \ldots A(s)}_{p+1 \text{ terms}} \right| \leq K \, \delta_1^{p+1} \ . \tag{4.8.1}$$

Proof. Denote by $\Pi(s, t; p)$ the product of matrices appearing in the left-hand side of (4.8.1). The proof goes along the same lines as (Moulines et al., 2005, Proposition 13) but we use the modulus of continuity instead of the β-Lipschitz norm to control the local oscillation of matrices.
For $\ell_1 \geq 1$ and any square matrices $A_1, \ldots, A_{\ell_1}$, adopting the convention $\prod_{i=i_1}^{i_2} A_i = A_{i_1} \ldots A_{i_2}$ if $i_1 \leq i_2$ and $\prod_{i=i_1}^{i_2} A_i$ is the identity matrix if $i_1 > i_2$, we have

$$\begin{aligned} \prod_{k=1}^{\ell_1} A_k &= A_1^{\ell_1} + \sum_{k=1}^{\ell_1 - 1} \left(A_1^{\ell_1 - k} \prod_{i=\ell_1 - k + 1}^{\ell_1} A_i - A_1^{\ell_1 - (k-1)} \prod_{i=\ell_1 - k + 2}^{\ell_1} A_i \right) \\ &= A_1^{\ell_1} + \sum_{k=1}^{\ell_1 - 1} A_1^{\ell_1 - k} \left(A_{\ell_1 - k + 1} - A_1 \right) \prod_{i=\ell_1 - k + 2}^{\ell_1} A_i \ . \end{aligned} \tag{4.8.2}$$

Given a positive integer ℓ, using the Euclidean division of $p + 1$ by ℓ, $p + 1 = \ell q + r$, we decompose the product $\Pi(s, t; p)$ as

$$\begin{aligned} \Pi(s, t; p) = \prod_{j=0}^{q-1} \left(\prod_{k=1}^{\ell} A\left(t - \frac{(j\ell + k - 1)(t-s)}{p} \right) \right) \\ \times \prod_{k=1}^{r} A\left(t - \frac{(q\ell + k - 1)(t-s)}{p} \right) \ . \end{aligned} \tag{4.8.3}$$

Using (4.8.2) we have for any $h \geq \ell(t-s)/p$, $0 \leq j \leq q$ and $0 \leq \ell_1 \leq \ell$,

$$\left| \prod_{k=1}^{\ell_1} A\left(t - \frac{(j\ell + k - 1)(t-s)}{p}\right) \right| \leq \left| \left(A\left(t - \frac{j\ell(t-s)}{p}\right)\right)^{\ell_1} \right| + (\ell_1 - 1)\,(A^*)^{\ell_1 - 1}\,\omega_h(A; I) \ . \quad (4.8.4)$$

Take an arbitrary $\delta_2 \in (\delta, \delta_1)$ (say the middle point). The eigenvalues of A are at most δ on I and $A^* < \infty$. Applying (Moulines et al., 2005, Lemma 12) we obtain that there is a constant $K_1 \geq 1$ only depending on δ, δ_2 and A^* such that $|\,(A(t - j\ell(t-s)/p))^{\ell_1}\,| \leq K_1 \delta_2^{\ell_1}$. From (4.8.3) and (4.8.4), we derive the following inequality:

$$|\Pi(s,t;p)| \leq \left(K_1\delta_2^{\ell} + K_2\omega_h(A;I)\right)^q \ \left(K_1\delta_2^{r} + K_2\omega_h(A;I)\right) \ .$$

where $K_2 = (\ell - 1)\,(\max\{A^*, 1\})^{\ell - 1}$.
We can choose a positive integer ℓ and a positive number ε_0 only depending on δ_2, δ_1 and K_1 such that

$$K_1\delta_2^{\ell} \leq \delta_1^{\ell} - \varepsilon_0 \ .$$

In the following, we set $\varepsilon = \varepsilon_0 / K_2$. The previous bound gives that for any $h \in (0,1)$ such that $\omega_h(A;I) \leq \varepsilon$ and $\ell(t-s)/p \leq h$,

$$|\Pi(s,t;p)| \leq \delta_1^{\ell q}\ (K_1\delta_2^{r} + \varepsilon_0) \leq K_1\delta_1^{p+1} + \varepsilon_0\delta_1^{\ell q} \leq \left(K_1 + \varepsilon_0 \max\left\{1, \delta_1^{1-\ell}\right\}\right)\ \delta_1^{p+1} \ .$$

Hence, we have the result. □

4.8.2 Proof of Proposition 2

We can now provide a proof of Proposition 2.
Equation (4.3.1) can be more compactly written as

$$X_{t,T} = \boldsymbol{\theta}'\left(\frac{t-1}{T}\right)\mathbf{X}_{t-1,T} + \sigma\left(\frac{t}{T}\right)\xi_{t,T} \ . \quad (4.8.5)$$

For all $k \geq 0$, iterating this recursive equation k times, we have

$$X_{t,T} = \boldsymbol{e}_1'\left[\prod_{i=1}^{k+1} A\left(\frac{t-i}{T}\right)\right]\boldsymbol{X}_{t-k-1,T} + \sum_{j=0}^{k} \sigma\left(\frac{t-j}{T}\right)\boldsymbol{e}_1'\left[\prod_{i=1}^{j} A\left(\frac{t-i}{T}\right)\right]\boldsymbol{e}_1\xi_{t-j} \ , \quad (4.8.6)$$

where $\boldsymbol{e}_1 = [1\ 0 \ldots 0]'$ and

$$A(u) = \begin{bmatrix} \theta_1(u) & \theta_2(u) & \ldots & \ldots & \theta_d(u) \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 0 \end{bmatrix} .$$

Note that the eigenvalues of $A(u)$ are the reciprocals of the roots of the local time varying autoregressive polynomial $z \mapsto \boldsymbol{\theta}(z;u)$, and thus are at most $\delta < 1$. Moreover, since $\boldsymbol{\theta}$ is bounded by a constant only depending on d and is uniformly continuous on $I = (-\infty, 1]$, so is A as a function defined on I and we can find $h \in (0,1)$ such that $\omega_h(A, I) \leq \varepsilon$ for any positive ε. If $\boldsymbol{\theta} \in \Lambda_d(\beta, R)$, this h can be chosen depending only on ε, β and R (and also on the matrix norm $|\cdot|$).
Consider $\delta_1 \in (\delta, 1)$. Lemma 13 gives that there exist some positive constant ε, ℓ and K only depending on A^*, δ and δ_1 such that, for any $h \in (0,1)$ fulfilling $\omega_h(A; I) \leq \varepsilon$, we have, for all $T \geq 1$, $t \leq T$ and $j \geq 1$ so that $T \geq \ell/h$,

$$\left| \prod_{i=1}^{j} A\left(\frac{t-i}{T}\right) \right| \leq K \, \delta_1^j \, .$$

We here consider the ℓ^∞ operator norm which is the maximum absolute row sum of the matrix, in which case $A^* = \max\{1, \sup_{u \in I}(|\theta_1(u)| + \cdots + |\theta_d(u)|)\} \leq 2^d d^{1/2}$. Hence, by (4.8.6) we obtain that

$$X_{t,T} = \sum_{i=1}^{d} b_{t,T}(k,i) X_{t-k-i,T} + \sum_{j=0}^{k} a_{t,T}(j) \, \sigma\left(\frac{t-j}{T}\right) \xi_{t-j,T}, \quad 1 \leq t \leq T \, . \tag{4.8.7}$$

with, provided that $T > \ell/h$, for all $t \leq T$, $k, j \geq 1$ and $i = 1, \ldots, d$,

$$\begin{aligned} \left| b_{t,T}(k,i) \right| &\leq K \delta_1^{k+1} \, , \\ \left| a_{t,T}(j) \right| &\leq K \delta_1^{j} \, . \end{aligned}$$

The result follows.

4.8.3 Proof of Lemma 6

The idea is to choose a convenient $i_N \in \{1, \ldots, N\}$ and use that

$$\min_{1 \leq i \leq N} S_T\left(\widehat{X}_T^{(\beta_i)}; \psi, \beta, R, \delta, \rho, \sigma_+\right) \leq S_T\left(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta, R, \delta, \rho, \sigma_+\right) \, .$$

We treat the cases $\beta_0 < \infty$ and $\beta_0 = \infty$ separately.
Let us first consider the case $\beta_0 < \infty$. Let $\beta \in (0, \beta_0)$, $\delta \in (0,1)$, $R > 0$ and $\rho \in [0,1]$. Let $i_N \in \{1, \ldots, N\}$ be such that $\beta_{i_N} = (i_N - 1)\beta_0/N < \beta \leq i_N \beta_0/N$. Since $C(\beta, R, \delta, \rho, \sigma_+) \subset C(\beta_{i_N}, R, \delta, \rho, \sigma_+)$, we have, for all $\delta \in (0,1)$, $R > 0$, $\rho > 0$ and $\sigma_+ > 0$,

$$\begin{aligned} T^{2\beta/(1+2\beta)} S_T\left(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta, R, \delta, \rho, \sigma_+\right) &\leq T^{2\beta/(1+2\beta)} S_T\left(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta_{i_N}, R, \delta, \rho, \sigma_+\right) \\ &\leq T^{2\beta_0/N} \, T^{2\beta_{i_N}/(1+2\beta_{i_N})} S_T\left(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta_{i_N}, R, \delta, \rho, \sigma_+\right) \, , \end{aligned}$$

where we used that $\beta_{i_N} < \beta \leq \beta_{i_N} + \beta_0/N$. Recall that we assumed $N \geq \lceil \log T \rceil$, so that $T^{2\beta_0/N} \leq \mathrm{e}^{2\beta_0}$. Now, since for N large enough β_{i_N} remains in a closed interval of $(0, \beta_0)$ we get by Definition 11 that

$$\limsup_{T \to \infty} T^{2\beta_{i_N}/(1+2\beta_{i_N})} S_T\left(\widehat{X}_T^{(\beta_{i_N})}; \psi, \beta_{i_N}, R, \delta, \rho, \sigma_+\right) < \infty \, ,$$

which concludes the proof in the case $\beta_0 < \infty$.
We next consider the case where $\beta_0 = \infty$. In this case we take i_N such that $\beta_{i_N} = (i_N - 1)/N^{1/2} < \beta \le i_N/N^{1/2}$ which defines $i_N \in \{1, \dots, N\}$ uniquely as soon as $N^{1/2} > \beta$. The remainder of the proof is similar to the case $\beta_0 < \infty$ using the bound

$$T^{2\beta/(1+2\beta)} \le T^{2/N^{1/2}}\, T^{2\beta_{i_N}/(1+2\beta_{i_N})} \le \mathrm{e}^2\, T^{2\beta_{i_N}/(1+2\beta_{i_N})}\,,$$

under the assumption $N \ge \lceil (\log T)^2 \rceil$.

4.8.4 Application to the TVAR process: proof of Theorem 4.3.2

Theorem 4.3.2 is an application of Theorem 4.2.1 to the aggregation of minimax predictors for the TVAR model (**M**-2).
We first note that Proposition 2 shows that, for T large enough the TVAR model (**M**-2) satisfies (**M**-1) with A_* bounded independently of T as in (4.3.5) and $Z_t = |\xi_t|$ for all $t \in \mathbb{Z}$. Hence Assumptions (**I**-1) and (**I**-2) respectively imply (**N**-1) and (**N**-2).
This shows that Theorem 4.2.1 applies under the assumptions of Theorem 4.3.2 and that the constants A_* and a^* appearing in (4.2.8), (4.2.9), (4.2.10) and (4.2.11) can be replaced by $\bar{K}\sigma_+/(1-\delta_1)$ and $\bar{K}\sigma_+$, respectively, where $\bar{K} > 0$ and $\delta_1 \in (0,1)$ can be chosen only depending on δ, β, and R.
On the other hand, Lemma 6 shows that, under the given assumptions on the predictors and with the given choices of N, the smallest prediction risk among the selected predictors, achieves a rate $T^{-2\beta/(1+2\beta)}$ for some positive constant C only depending on $\beta, \delta, R > 0, \rho$ and ψ. Hence, we get with Theorem 4.2.1 that

$$\limsup_{T\to\infty} T^{2\beta/(1+2\beta)} S_T\left(\widehat{X}_T; \psi, \beta, R, \delta, \rho, \sigma_+\right) \le C + \limsup_{T\to\infty} T^{2\beta/(1+2\beta)} \mathcal{R}(N,T)\,, \tag{4.8.8}$$

where C is a positive constant and $\mathcal{R}(N,T)$ is a remaining term which, in the setting (i) in Theorem 4.3.2, is given by

$$\mathcal{R}(N,T) = \frac{\log N}{T\eta} + 2\eta\,(1+L_*)^4\, m_4\, \frac{\bar{K}^4\sigma_+^4}{(1-\delta_1)^4}\,, \tag{4.8.9}$$

in the setting (ii), is given by

$$\mathcal{R}(N,T) = \frac{\log N}{T\eta} + (2\eta)^{p/2-1}(1+L_*)^p m_p\, \frac{\bar{K}^p\sigma_+^p}{(1-\delta_1)^p}\,, \tag{4.8.10}$$

and, in the setting (iii), taking $\lambda = \zeta/(\bar{K}\sigma_+(L_*+1))$, is given by

$$\mathcal{R}(N,T) = \frac{\log N}{T\eta} + \frac{2}{\mathrm{e}}\,(\phi(\zeta))^{1/(1-\delta_1)}\,\lambda^{-2}\left(2+\lambda(2\eta)^{-1/2}\right)\mathrm{e}^{-\lambda(2\eta)^{-1/2}}\,. \tag{4.8.11}$$

Replacing η and N in (4.8.9) as given by (i) and (4.3.11), we get

$$\sigma_+^{-2}\,\mathcal{R}(N,T) \le \left(\frac{\log\lceil \log T\rceil}{T}\right)^{1/2}\left(1 + 2\,(1+L_*)^4\, m_4\, \frac{\bar{K}^4}{(1-\delta_1)^4}\right).$$

Hence, using that $\beta < \beta_0 \leq 1/2$, this upper bound is negligible with respect to $T^{-2\beta/(2\beta+1)}$ and, with (4.8.8), we get (4.3.12).
Analogously, we replace η and N in (4.8.10) as given by (ii) and (4.3.11), we get

$$\sigma_+^{-2}\,\mathcal{R}(N,T) \leq \left(\frac{\log\lceil \log T\rceil}{T}\right)^{1-2/p}\left(1+2^{p/2-1}\,(1+L_*)^p\,m_p\,\frac{\bar{K}^p}{(1-\delta_1)^p}\right).$$

Since $\beta < \beta_0 \leq (p-2)/4$, this upper bound is negligible with respect to $T^{-2\beta/(2\beta+1)}$ and, with (4.8.8), we get (4.3.12).
Finally, in the setting (iii), using the specific form of η, we get from (4.8.11) that

$$\sigma_+^{-2}\,\mathcal{R}(N,T) \leq \frac{1}{T}\left[(\log T)^3\log\left(\lceil\log T\rceil^2\right)+c_1\left(1+(\log T)^{3/2}\right)T^{-c_2(\log T)^{1/2}}\right],$$

where c_1 and c_2 are positive constants only depending on ζ, $\phi(\zeta)$, δ_1, $\bar{K}$ and L_*. For any $\beta > 0$, this upper bound is negligible with respect to $T^{-2\beta/(2\beta+1)}$ and, with (4.8.8), we get (4.3.12).

4

4.8.5 Proof of Lemma 7

Denote $\omega(x) = \min\{2^{-1/2}, \max\{x, -2^{-1/2}\}\}$, so that $\omega(x)^2 = \min(1/2, x^2) \leq x^2$. The function $x \mapsto \exp(-x^2)$ is concave on $[-2^{-1/2}, 2^{-1/2}]$, so introducing $\omega(x)$ and then using Jensen's inequality, we get

$$\begin{aligned}\int \exp\left(-x^2\right)\mathrm{d}\mathbb{P}(x) &\leq \int \exp\left(-\omega^2(x)\right)\mathrm{d}\mathbb{P}(x) \leq \exp\left(-\left(\int \omega(x)\,\mathrm{d}\mathbb{P}(x)\right)^2\right)\\ &= \exp\left(-\left(\int x\mathrm{d}\mathbb{P}(x)\right)^2+\left(\int x\mathrm{d}\mathbb{P}(x)\right)^2-\left(\int \omega(x)\,\mathrm{d}\mathbb{P}(x)\right)^2\right).\end{aligned}$$

It only remains to show that $(\int x\mathrm{d}\mathbb{P}(x))^2 - (\int \omega(x)\mathrm{d}\mathbb{P}(x))^2 \leq (a^2-1/2)_+$, with the assumption that $\mathbb{P}$ has support on $[-a,a]$. This is verified if $a \leq 2^{-1/2}$, so we now assume $a > 2^{-1/2}$. We write

$$\left(\int x\mathrm{d}\mathbb{P}(x)\right)^2-\left(\int \omega(x)\,\mathrm{d}\mathbb{P}(x)\right)^2 = \int (x-\omega(x))\,(y+\omega(y))\;\mathrm{d}\mathbb{P}(x)\,\mathrm{d}\mathbb{P}(y).$$

We note that $|x-\omega(x)| = (|x|-1/2)_+$ and $|y+\omega(y)| \in \{2|y|, |y|+2^{-1/2}\}$. We deduce that the product $(x-\omega(x))(y+\omega(y))$ either take non-positive values or positive values of the form

$$\begin{cases} 2\,|y|\left(|x|-2^{-1/2}\right) & \text{with } |x|>2^{-1/2}, |y|<2^{-1/2},\\ \left(|x|-2^{-1/2}\right)\left(|y|+2^{-1/2}\right) & \text{with } |x|>2^{-1/2}, |y|>2^{-1/2}.\end{cases}$$

Now, for $x, y \in [-a,a]$ with $a > 2^{-1/2}$, in the first case, we have $2|y|(|x|-2^{-1/2}) \leq 2^{1/2}(a-2^{-1/2}) \leq a^2-1/2$ since $2^{1/2} \leq a+2^{-1/2}$, and, in the second case, $(|x|-2^{-1/2})(|y|+2^{-1/2}) \leq (a-2^{-1/2})(a+2^{-1/2}) = a^2-1/2$. The lemma follows.

4.8.6 Proof of Lemma 8

We define $\hat{\mathfrak{e}}$ as the (random) smallest index which minimizes $d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_j\})$ over $j \in \{0, \dots, M\}$ so that $d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_{\hat{\mathfrak{e}}}\}) = \min_{\theta \in \{\theta_0, \dots, \theta_M\}} d_X(\widehat{\theta}, \mathrm{v}_T\{\theta\})$. Note that $d_{X,T}(\theta_{\hat{\mathfrak{e}}}, \theta_j) \le d_X(\mathrm{v}_T\{\theta_{\hat{\mathfrak{e}}}\}, \widehat{\theta}) + d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_j\}) \le 2 d_X(\widehat{\theta}, \mathrm{v}_T\{\theta_j\})$. Hence

$$\begin{aligned}\max_{j=0,\dots,M} \mathbb{E}_{\theta_j}\left[d_X^2(\widehat{\theta}, \mathrm{v}_T\{\theta_j\})\right] &\ge \frac{1}{4} \max_{j=0,\dots,M} \mathbb{E}_{\theta_j}\left[d_{X,T}^2(\theta_{\hat{\mathfrak{e}}}, \theta_j)\right] \\ &\ge \frac{r^2}{4} \max_{j=0,\dots,M} \mathbb{P}_{\theta_j}\left(\{\hat{\mathfrak{e}} \neq j\} \cap \left\{\min_{i:i\neq j} d_{X,T}(\theta_i, \theta_j) > r\right\}\right) \\ &\ge \frac{r^2}{4}\left(1 - \min_{j=0,\dots,M} \mathbb{P}_{\theta_j}(\hat{\mathfrak{e}} = j) - \max_{j=0,\dots,M} \mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} d_{X,T}(\theta_i, \theta_j) \le r\right)\right).\end{aligned}$$

Birgé's lemma (Massart, 2007, Corollary 2.18) implies that

$$\min_{j=0,\dots,M} \mathbb{P}_{\theta_j}(\hat{\mathfrak{e}} = j) \le \max\left\{\left(\frac{2\mathrm{e}}{2\mathrm{e}+1}\right), \left(\frac{\max_{j=0,\dots,M} \mathcal{K}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0})}{\log(1+M)}\right)\right\},$$

so the lemma follows from Condition (4.5.5).

4.8.7 Proof of Lemma 9

By (4.5.11), we have $\theta_j \in s_1(\delta)$ for all $j = 0, \dots, M$. Decompose the Hölder-exponent $\beta = k + \alpha$ where k is an integer and $\alpha \in (0, 1]$. Differentiating (4.5.7) k times, we have, as in (4.5.10),

$$\theta_j^{(k)}(x) = \frac{R_0}{m^\alpha}\, w_{\lfloor mx \rfloor + 1}^{(j)}\, K^{(k)}\left(\{mx\} - \frac{1}{2}\right), \quad \text{for all } x \in [0, 1].$$

Thus, for s, s' in the same interval $[\ell/m, (\ell+1)/m]$ with $\ell = 0, \dots, m-1$, we get

$$\begin{aligned}\left|\theta_j^{(k)}(s) - \theta_j^{(k)}(s')\right| &\le \frac{R_0}{m^\alpha}\left|K^{(k)}\left(m\, s - \ell - \frac{1}{2}\right) - K^{(k)}\left(m\, s' - \ell - \frac{1}{2}\right)\right| \\ &\le R_0\ |K|_\beta\ |s - s'|^\alpha\end{aligned}$$

The same inequality then follows with R_0 replaced by $2R_0$ for s, s' in two such consecutive intervals. Now, if s, s' are separated by at least one such interval, we have $|s - s'| \ge m^{-1}$ and, using that K has support in $(-1/2, 1/2)$, we have that $|K^{(k)}(x)|$ is bounded by $|K|_\beta$. We thus get in this case that

$$\left|\theta_j^{(k)}(s) - \theta_j^{(k)}(s')\right| \le \frac{2R_0}{m^\alpha} \sup_{-1/2 \le x \le 1/2} \left|K^{(k)}(x)\right| \le 2R_0\ |K|_\beta\ |s - s'|^\alpha\ .$$

The last two displays and (4.5.8) then yields $\theta_j \in \Lambda_1(\beta, R)$.

4.8.8 Proof of Lemma 10

Let $j = 1, \dots, M$. Recall that $\theta_0 \equiv 0$ by (4.5.6) and (4.5.7). By (4.5.9) and (4.5.1), we have that $(X_{s,T})_{s\le 0}$ has the same distribution under $\mathbb{P}_{\theta_j}$ and $\mathbb{P}_{\theta_0}$ (which is the distribution of $(\xi_s)_{s\le 0}$). Hence, the likelihood ratio $\mathrm{d}\mathbb{P}_{\theta_j}/\mathrm{d}\mathbb{P}_{\theta_0}$ of $(X_{s,T})_{s\le T}$ is given by the corresponding conditional likelihood ratio of $(X_{s,T})_{1\le s\le T}$ given $(X_{s,T})_{s\le 0}$. Hence, under (I-3), we obtain that

$$\frac{\mathrm{d}\mathbb{P}_{\theta_j}}{\mathrm{d}\mathbb{P}_{\theta_0}} = \prod_{t=1}^{T} \frac{f\left(X_{t,T} - \theta_j((t-1)/T)X_{t-1,T}\right)}{f\left(X_{t,T} - \theta_0((t-1)/T)X_{t-1,T}\right)} = \prod_{t=1}^{T} \frac{f\left(X_{t,T} - \theta_j((t-1)/T)X_{t-1,T}\right)}{f\left(X_{t,T}\right)} ,$$

where, in the second equality, we used again that $\theta_0 \equiv 0$. Now, under $\mathbb{P}_{\theta_j}$, we have $X_{t,T} = \theta_j((t-1)/T)X_{t-1,T} + \xi_t$. Thus, we get

$$\begin{aligned}
\mathcal{K}\left(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}\right) &= \mathbb{E}_{\theta_j}\left[\log \frac{\mathrm{d}\mathbb{P}_{\theta_j}}{\mathrm{d}\mathbb{P}_{\theta_0}}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\theta_j}\left[\log \frac{f(\xi_t)}{f(\theta_j((t-1)/T)X_{t-1,T} + \xi_t)}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\theta_j} \int \log\left(\frac{f(u)}{f(\theta_j((t-1)/T)X_{t-1,T} + u)}\right) f(u)\,\mathrm{d}u
\end{aligned}$$

Using Assumption (I-3) yields

$$\mathcal{K}\left(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}\right) \le \sum_{t=1}^{T} \mathbb{E}_{\theta_j}\left[\kappa\theta_j^2\left(\frac{t-1}{T}\right)X_{t-1,T}^2\right] \le \kappa\theta^{*2} \sum_{t=1}^{T} \mathbb{E}_{\theta_j}\left[X_{t-1,T}^2\right]. \tag{4.8.12}$$

The series representation (4.5.2), the fact that ξ is centered with unit variance and (4.5.11) imply that for all $t = 0, \dots, T$

$$\mathbb{E}_{\theta_j}\left[X_{t,T}^2\right] \le \left(1 - \theta^{*2}\right)^{-1} .$$

Using this bound and (4.5.11) in (4.8.12), we obtain

$$\mathcal{K}\left(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_0}\right) \le \frac{R_0^2\, \mathrm{e}^{-2}\, \kappa\, T}{(1-\delta^2)\, m^{2\beta}} .$$

The proof of Lemma 10 now follows by applying the first bound in (4.5.6).

4.8.9 Proof of Lemma 11

The proof relies on an upper bound of $d_{X,T}^2(\theta_i, \theta_j)$ involving the noise (ξ_t). By the expression of θ_j in (4.5.10), we have

$$d_{X,T}^2(\theta_i, \theta_j) = \frac{R_0^2}{T m^{2\beta}} \sum_{t=0}^{T-1} X_{t,T}^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right) , \tag{4.8.13}$$

where we denoted $\varphi(t) = \{mt/T\} - 1/2$ and $k(t) = \lfloor mt/T \rfloor + 1$. Using (4.5.2) and (4.5.11), we have, for all $0 \le t \le T-1$,

$$\left|X_{t,T}\right| \ge |\xi_t| - \sum_{j=1}^{\infty} \theta^{*j} \left|\xi_{t-j}\right| \ ,$$

which implies

$$X_{t,T}^2 \ge \xi_t^2 - 2\,|\xi_t| \sum_{j=1}^{\infty} \theta^{*j} \left|\xi_{t-j}\right| \ .$$

Inserting this bound in (4.8.13), we get

$$\frac{m^{2\beta}}{R_0^2}\, d_{X,T}^2(\theta_i, \theta_j) \ge \frac{1}{T} \sum_{t=0}^{T-1} \xi_t^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right) - \mathcal{R}_T \ , \tag{4.8.14}$$

where

$$\mathcal{R}_T = \frac{2\mathrm{e}^{-2}}{T} \sum_{t=0}^{T-1} \sum_{j=1}^{\infty} \theta^{*j}\, |\xi_t| \, \left|\xi_{t-j}\right|$$

Thus, with (4.8.14), the left-hand side of inequality (4.5.12) is upper bounded by

$$\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} \frac{1}{T} \sum_{t=0}^{T-1} \xi_t^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right) < 2A\right) + \mathbb{P}(\mathcal{R}_T > A) \ .$$

Using that ξ is centered with unit variance and then (4.5.11), we easily get that

$$\mathbb{E}_{\theta_j}\left[\mathcal{R}_T\right] \le \frac{2\mathrm{e}^{-2}}{T} \sum_{t=0}^{T-1} \sum_{j=1}^{\infty} \theta^{*j} \le \frac{2\mathrm{e}^{-2}\theta^*}{1-\theta^*} \le \frac{2R_0\mathrm{e}^{-3}}{(1-\delta)m^{\beta}} \ .$$

Hence, by Markov's inequality, to conclude the proof, it now suffices to show that, for A well chosen,

$$\mathbb{P}_{\theta_j}\left(\min_{i:i\neq j} \frac{1}{T} \sum_{t=0}^{T-1} \xi_t^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)}\right)^2 K^2\left(\varphi\left(t\right)\right) < 2A\right) \le \varepsilon \ . \tag{4.8.15}$$

For $k \in \{1, \ldots, m\}$ we define $J_k = \{\lfloor (k-1)T/m \rfloor + i \ : \ \lceil T/(4m) \rceil + 1 \le i \le \lfloor 3T/(4m) \rfloor\}$. We observe that the cardinality of J_k is

$$\Gamma\left(\frac{T}{m}\right) = \left\lfloor \frac{3T}{4m} \right\rfloor - \left\lceil \frac{T}{4m} \right\rceil \ge 1 \ ,$$

where the lower bound is a consequence of the assumption $T \ge 4m$ in the lemma. Moreover, it is easy to check that we have $|\varphi(t)| \le 1/4$ for all index $t \in J_k$ and that,

for each $1 \le k \le m$, the set J_k is included in the set $\{1 \le t \le T-1 : k(t) = k\}$ (so that, in particular, $J_k \cap J_{k'} = \emptyset$ for $k < k'$). It follows that random variables

$$S_k = \frac{1}{\Gamma(T/m)} \sum_{t \in J_k} \xi_{t-1}^2, \quad \text{for } k = 1, \dots, m$$

are i.i.d. By the monotonicity of K in $\mathbb{R}_-$ and its symmetry we have

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \xi_t^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)} \right)^2 K^2(\varphi(t)) &\ge \frac{1}{T} \sum_{k=1}^{m} \left(w_k^{(i)} - w_k^{(j)} \right)^2 \sum_{t \in J_k} \xi_t^2 K^2(\varphi(t)) \\ &\ge \frac{K^2(1/4)\Gamma(T/m)}{T} \sum_{k=1}^{m} \left(w_k^{(i)} - w_k^{(j)} \right)^2 S_k. \end{aligned}$$

From (4.5.6), for any $i, j \in \{1, \dots, M\}$ there exist at least $\lceil m/8 \rceil$ values of k for which $(w_k^{(i)} - w_k^{(j)})^2$ equals one in the above sum. Hence using the order statistics $S_{(1,m)} \le \dots \le S_{(m,m)}$, we thus obtain that

$$\begin{aligned} \min_{i:i \ne j} \frac{1}{T} \sum_{t=0}^{T-1} \xi_t^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)} \right)^2 K^2(\varphi(t)) &\ge \frac{K^2(1/4)\Gamma(T/m)}{T} \sum_{k=1}^{\lceil m/8 \rceil} S_{(k,m)} \\ &\ge \frac{K^2(1/4)\, m\, \Gamma(T/m)}{16\, T} S_{(\lfloor m/16 \rfloor, m)} \\ &\ge \frac{K^2(1/4)}{128} S_{(\lfloor m/16 \rfloor, m)}, \end{aligned}$$

where we used $\Gamma(T/m) \ge T/(8m)$ for $T/m \ge 4$ in the last inequality. Let us denote by F the cumulative distribution function of S_1, which only depends on $\Gamma(T/m)$ and on the distribution of ξ_0. For $x > 0$, we have

$$\begin{aligned} \mathbb{P}\left(S_{(\lfloor m/16 \rfloor, m)} \le x \right) &= \mathbb{P}\left(\mathrm{Bin}(m, F(x)) \ge \left\lfloor \frac{m}{16} \right\rfloor \right) \\ &\le \frac{m}{\lfloor m/16 \rfloor} F(x) \le 32 F(x). \end{aligned}$$

Gathering the last two bounds, we get that

$$\begin{aligned} \mathbb{P}_{\theta_j} \left(\min_{i:i \ne j} \frac{1}{T} \sum_{t=1}^{T-1} \xi_t^2 \left(w_{k(t)}^{(i)} - w_{k(t)}^{(j)} \right)^2 K^2(\varphi(t)) \le 2A \right) &\le \mathbb{P}\left(S_{(\lfloor m/16 \rfloor, m)} \le \frac{256\, A}{K^2(1/4)} \right) \\ &\le 32\, F\left(\frac{256\, A}{K^2(1/4)} \right). \end{aligned}$$

Recall that $\Gamma(T/m) \ge 1$ and note that S_1 admits a density, since ξ does. By the strong law of large numbers, we further have that the random variable S_1 converges to 1 almost surely when $\Gamma(T/m)$ goes to infinity, so there exists $x_0 > 0$ depending only on the density of ξ such that $F(x_0) \le \varepsilon/32$ whatever the value of $\Gamma(T/m) \ge 1$. Therefore, there exists some $A > 0$, depending only on the distribution of ξ, such that (4.8.15) holds, which achieves the proof.

4.9 From best predictor regret bounds to convex regret bounds

We can improve upon the convex regret bound (4.2.15) when T is larger than N^2 and the noise Z either satisfies (N-2) for some positive ζ, or (N-1) with $p > 4$. The improvement is based on the following deterministic lemma adapted from the proof of Theorem 6 in Yang (2004).

Lemma 14. *Let $(x_t)_{1\le t\le T}$ be a real valued sequence and $\{(\widehat{x}_t^{(i)})_{1\le t\le T}, 1 \le i \le N\}$ be a collection of predicting sequences. For any $\alpha \in \mathcal{S}_N$ (defined by (4.2.4)), we set $\widehat{x}_t^{[\alpha]} = \sum_{i=1}^N \alpha_i \widehat{x}_t^{(i)}$. For any $T \ge N$ and $\eta > 0$, there exists an aggregated predictor $(\widehat{x}_t)_{1\le t\le T}$ such that,*

$$\frac{1}{T}\sum_{t=1}^T (\widehat{x}_t - x_t)^2 \le \inf_{\nu\in\mathcal{S}_N} \frac{1}{T}\sum_{t=1}^T \left(\widehat{x}_t^{[\nu]} - x_t\right)^2 + \frac{N}{\eta T}\log\left(\frac{5T}{N}\right) + \frac{3N}{T}\times\frac{1}{T}\sum_{t=1}^T y_t^2 + \frac{1}{T}\sum_{t=1}^T \left(y_t^2 - \frac{1}{2\eta}\right)_+ , \quad (4.9.1)$$

with $y_t = |x_t| + \max_{1\le i\le N} |\widehat{x}_t^{(i)}|$.

Proof. We set $\varepsilon = N/T \le 1$. Let $\mathcal{S}_{N,\varepsilon} \subset \mathcal{S}_N$ be a minimal ε-net of the simplex $\mathcal{S}_N$ for the ℓ_1 distance. We consider the aggregated predictor $(\widehat{x}_t)_{1\le t\le T}$ obtained from the collection of predictors $\{(\widehat{x}_t^{[\alpha]})_{1\le t\le T}, \alpha \in \mathcal{S}_{N,\varepsilon}\}$ with the weights (4.2.6). From Lemma 5, we have

$$\frac{1}{T}\sum_{t=1}^T (\widehat{x}_t - x_t)^2 \le \min_{\alpha'\in\mathcal{S}_{N,\varepsilon}} \frac{1}{T}\sum_{t=1}^T \left(\widehat{x}_t^{[\alpha']} - x_t\right)^2 + \frac{\log N_\varepsilon}{T\eta} + \frac{1}{T}\sum_{t=1}^T \left(y_{t,\varepsilon}^2 - \frac{1}{2\eta}\right)_+ , \quad (4.9.2)$$

where $N_\epsilon = |\mathcal{S}_{N,\varepsilon}|$ and $y_{t,\varepsilon} = |x_t| + \max_{\alpha'\in\mathcal{S}_{N,\varepsilon}} |\widehat{x}_t^{[\alpha']}|$.
Note that, for any $\alpha, \alpha' \in \mathcal{S}_N$, we have

$$|\widehat{x}_t^{[\alpha']} - \widehat{x}_t^{[\alpha]}| \le \sum_{j=1}^N |\alpha_j - \alpha'_j| \max_{i=1,\dots,N} |\widehat{x}_t^{(i)}| .$$

Hence, for any $\alpha \in \mathcal{S}_N$ there exists $\alpha' \in \mathcal{S}_{N,\varepsilon}$ such that $|\widehat{x}_t^{[\alpha']} - \widehat{x}_t^{[\alpha]}| \le \varepsilon y_t$ and

$$\left(\widehat{x}_t^{[\alpha']} - x_t\right)^2 = \left(\widehat{x}_t^{[\alpha']} - \widehat{x}_t^{[\alpha]}\right)^2 + 2\left(\widehat{x}_t^{[\alpha']} - \widehat{x}_t^{[\alpha]}\right)\left(\widehat{x}_t^{[\alpha]} - x_t\right) + \left(\widehat{x}_t^{[\alpha]} - x_t\right)^2 \le (\varepsilon^2 + 2\varepsilon)y_t^2 + \left(\widehat{x}_t^{[\alpha]} - x_t\right)^2 .$$

Plugging this bound in (4.9.2) and using that, since $\mathcal{S}_{N,\varepsilon} \subset \mathcal{S}_N$, $y_{t,\epsilon} \le y_t$, we obtain

$$\frac{1}{T}\sum_{t=1}^T (\widehat{x}_t - x_t)^2 \le \inf_{\alpha\in\mathcal{S}_N} \frac{1}{T}\sum_{t=1}^T \left(\widehat{x}_t^{[\alpha]} - x_t\right)^2 + \frac{\log N_\varepsilon}{T\eta} + \frac{3\varepsilon}{T}\sum_{t=1}^T y_t^2 + \frac{1}{T}\sum_{t=1}^T \left(y_t^2 - \frac{1}{2\eta}\right)_+ . \quad (4.9.3)$$

For any $0 < \varepsilon \leq 1$, the cardinality N_ε of a minimal ε-net of $\mathcal{S}_N$ can be upper-bounded by $(5/\varepsilon)^N$, see (Ghosal and van der Vaart, 2001, Lemma A.4). So, for the choice $\varepsilon = N/T \leq 1$, we get the bound (4.9.1). □

We can now investigate how the bound (4.2.15) can be improved when conditions stronger than (N-1) with $p = 4$ are imposed on the noise Z.

Theorem 4.9.1. *Assume that Assumption (M-1) holds and let $\{(\widehat{X}_t^{(i)})_{1\leq t\leq T},\ 1 \leq i \leq N\}$ be a collection of sequences of L-Lipschitz predictors with L satisfying (L-1).*

(i) Assume that the noise Z satisfies (N-1) with a given $p > 2$ and let $\widehat{X} = (\widehat{X}_t)_{1\leq t\leq T}$ denote the aggregated predictor defined in the proof of Lemma 14 with

$$\eta = \frac{1}{2m_p^{2/p}(1+L_*)^2 A_*^2}\left(\frac{N}{T}\log\left(\frac{5T}{N}\right)\right)^{2/p}. \tag{4.9.4}$$

For $T \geq N$, we then have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \min_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(i)} - X_t\right)^2\right] + C_1\left(\frac{N}{T}\log\left(\frac{5T}{N}\right)\right)^{1-2/p}, \tag{4.9.5}$$

with $C_1 = 6m_p^{2/p}(1+L_*)^2 A_*^2$.

(ii) Assume that the noise Z fulfills (N-2) for some positive ζ and let $\widehat{X} = (\widehat{X}_t)_{1\leq t\leq T}$ denote the aggregated predictor obtained using the weights (4.2.6) with

$$\eta = \frac{\zeta^2}{2A_*^2(L_*+1)^2}\left(\log\left(\frac{T}{\log N}\right)\right)^{-2}. \tag{4.9.6}$$

For $T \geq N$, we then have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{X}_t^{[\nu]} - X_t\right)^2 + \frac{2A_*^2(L_*+1)^2}{\zeta^2}\frac{N}{T}\left\{\left(\log\left(\frac{T}{\log N}\right)\right)^2\log\left(\frac{5T}{N}\right) + \frac{\phi(\zeta)}{\mathrm{e}}\left(7+\log\left(\frac{T}{\log N}\right)\right)\right\}. \tag{4.9.7}$$

(Note that when $N/T \to 0$, the term between curly brackets is equivalent to $(\log T)^3$).

Proof. We define $Y_t = |X_t| + \max_{1\leq i\leq N}|\widehat{X}_t^{(i)}|$.
Case (i). Following the same lines as in the proof of Theorem 4.2.1, we obtain that $\mathbb{E}[Y_t^2] \leq A_*^2(1+L_*)^2 m_2$ and $\mathbb{E}[(Y_t^2 - 1/(2\eta))_+] \leq (2\eta)^{p/2-1}A_*^p(1+L_*)^p m_p$. Hence, from

Lemma 14 we get

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{X}_t - X_t\right)^2 \le \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{X}_t^{[\nu]} - X_t\right)^2 + \frac{3N}{T}A_*^2(1+L_*)^2 m_2 + \frac{N}{\eta T}\log\left(\frac{5T}{N}\right) + (2\eta)^{p/2-1}A_*^p(1+L_*)^p m_p\,. \quad (4.9.8)$$

Since $m_2 \le m_p^{2/p}$, for η given by (4.9.4), the inequality (4.9.5) follows.

Case (ii). We set $\lambda = \zeta/(A_*(1+L_*))$. Following the same lines as in the proof of Theorem 4.2.1, we obtain that $\mathbb{E}[Y_t^2] \le \lambda^{-2}\phi(\zeta)$ and $\mathbb{E}[(Y_t^2 - 1/(2\eta))_+] \le 2\mathrm{e}^{-1}\lambda^{-2}(2+\lambda(2\eta)^{-1/2})\,\mathrm{e}^{-\lambda(2\eta)^{-1/2}}\phi(\zeta)$. Hence, from Lemma 14 we get that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \inf_{\nu\in\mathcal{S}_N}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{X}_t^{[\nu]} - X_t\right)^2 + \frac{N}{\eta T}\log\left(\frac{5T}{N}\right) + \frac{3N}{\lambda^2 T}\phi(\zeta) + \frac{2}{\mathrm{e}}\lambda^{-2}\left(2+\lambda(2\eta)^{-1/2}\right)\mathrm{e}^{-\lambda(2\eta)^{-1/2}}\phi(\zeta)\,. \quad (4.9.9)$$

Choosing η as in (4.9.6), we obtain (4.9.7). □

Remark 10. For $p > 4$, we observe that the bound (4.9.5) improves upon (4.2.15) when $N\log(T/N) \le T^{(p-4)/(2p-4)}$. Similarly, the bound (4.9.7) improves upon (4.2.15) when $T \ge N^2(\log T)^6$.

Remark 11. The cardinality of a N/T-net of $\mathcal{S}_N$ roughly scales as $(T/N)^{N-1}$ with T, so the computational cost of the aggregated predictor $\widehat{X}$ of Lemma 14 is prohibitive. Hence, the bounds (4.9.5) and (4.9.7) are of theoretical interest only.

ACKNOWLEDGEMENTS

We gratefully acknowledge the fruitful comments of the referees. This work has been partially supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île de France (RDM-IdF) for the period 2012-2015 and by the Labex LMH (ANR-11-IDEX-003-02).

5 Locally stationary processes prediction by auto-regression

Abstract

In this contribution we introduce locally stationary time series through the local approximation of the non-stationary covariance structure by a stationary one. This allows us to define autoregression coefficients in a non-stationary context, which, in the particular case of a locally stationary Time Varying Autoregressive (TVAR) process, coincide with the generating coefficients. We provide and study an estimator of the time varying autoregression coefficients in a general setting. The proposed estimator of these coefficients enjoys an optimal minimax convergence rate under limited smoothness conditions. In a second step, using a bias reduction technique, we derive a minimax-rate estimator for arbitrarily smooth time-evolving coefficients, which outperforms the previous one for large data sets. For TVAR, the predictor naturally obtained from the estimator also exhibits an optimal minimax convergence rate.

5.1 Introduction

In many applications, one is interested in predicting the next values of an observed time series. It is the case in various areas like finance (stock market, volatility on prices), social sciences (population studies), epidemiology, meteorology and network systems (Internet traffic). Autoregressive processes have been used successfully in a stationary context for several decades. On the other hand, in a context where the number of observations can be very large, the usual stationarity assumption has to be weakened to take into account some smooth evolution of the environment.

Many prediction methods developed in signal processing are well known to adapt to a changing environment. This is the case of the wide spread recursive least square algorithms. The initial goal of these methods is to provide an online algorithm for estimating a regression vector with low numerical cost. Such methods usually rely on a forgetting factor or a gradient step size γ and they can be shown to be consistent in a stationary environment when γ decreases adequately to zero (see e.g. Duflo (1997)). However when the environment is changing, that is, when the regression parameter evolves along the time, a "small enough" γ often yields a good tracking of the evolving regression parameter. In order to have a sound and comprehensive understanding of this phenomenon, an interesting approach is to consider a local stationarity assumption,

as successfully initiated in Dahlhaus (1996b) by relying on a non-stationary spectral representation introduced in Priestley (1965); see also Dahlhaus (2012) and the references therein for a recent overview. The basic idea is to provide an asymptotic analysis for the statistical inference of non-stationary time series such as time varying autoregressive (TVAR) processes by relying on local stationary approximations. The analysis of the NLMS algorithm for tracking a moving autoregression parameter in this framework is tackled in Moulines et al. (2005). Such an analysis is based on the usual tools of non-parametric statistics. The TVAR parameter $\boldsymbol{\theta}$ is seen as the regular samples of a smooth $\mathbb{R}^d$-valued function. An in-fill asymptotic allows one to derive minimax rates of convergence for estimating this function on a fixed interval $[0, 1]$ within particular smoothness classes of functions. As shown in Moulines et al. (2005), it turns out that the NLMS algorithm provides an optimal minimax rate for estimating the TVAR parameters with Hölder smoothness index $\beta \in (0, 1]$ but is no longer optimal for $\beta > 1$, that is when the TVAR parameters are smoother than a continuously differentiable function. An improvement of the NLMS is proposed in Moulines et al. (2005) to cope with the case $\beta \in (0, 2]$ but, to the best of our knowledge, there is no available method neither for the $\boldsymbol{\theta}$ minimax-rate estimation nor for the minimax-rate prediction when $\beta > 2$, that is when the TVAR parameters are smoother than a two-times continuously differentiable function.

In the present work, our main contribution is twofold. First we introduce the concept of time-varying linear prediction coefficients to a general class of locally stationary processes. This general class extends the class of locally stationary processes as introduced in Dahlhaus (1996b) in a way that we believe is more natural and appropriate to the signal processing community. In the specific case of a TVAR process, these coefficients correspond to the time-varying autoregression parameters. Second, we show that the Yule-Walker estimator introduced in Dahlhaus and Giraitis (1998) for TVAR processes also applies to this general class and is minimax-rate for Hölder index $\beta = 2$. Moreover, by applying a bias reduction technique, we derive a new estimator which is minimax-rate for any Hölder index $\beta \geq 2$.

The paper is organized as follows. In Section 5.2, we introduce the locally stationary time series and define the regression problem investigated in this work. The Yule-Walker estimator is detailed in Section 5.4. In Section 5.3, we explain why and how minimax estimation is crucial for deriving practical predictors. Main results are presented in Section 5.5 relying on Hölder smoothness assumptions on the local spectral density of the locally stationary time series. The particular case of TVAR processes is treated in Section 5.6. Numerical experiments complete our study in Section 5.7, confirming the benefits of our approach when the length of the data set becomes very large.

Four appendices complete this paper. Section 5.8 contains useful results on locally stationary time series needed for showing the main theorems of Section 5.5. The proof of the main theorems of Section 5.5 are provided in Section 5.9. Some useful technical results can be found in Section 5.10. As a support of Section 5.8, we refer to the basic tool-kit on weakly stationary processes presented in Section 5.11.

5.2 General setting

In the following, vectors are denoted using boldface symbols, $\|\mathbf{x}\|$ denotes the Euclidean norm of $\mathbf{x}$, $\|\mathbf{x}\| = (\sum_i |x_i|^2)^{1/2}$, and $\|\mathbf{x}\|_1$ its ℓ_1 norm, $\|\mathbf{x}\|_1 = \sum_i |x_i|$. If f is a function, $\|f\|_\infty = \sup_x |f(x)|$ corresponds to its sup norm.

5.2.1 Main definitions

We consider a doubly indexed time series $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$, which we assume to be centred for convenience. Here t refers to a discrete time index and T is an additional index indicating the sharpness of the *local approximation* of the time series $(X_{t,T})_{t\in\mathbb{Z}}$ by a stationary one. Coarsely speaking, $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ is considered to be *locally stationary* if, for T large, given a set S_T of sample indices such that $t/T \approx u$ over $t \in S_T$, the sample $(X_{t,T})_{t\in S_T}$ can be approximately viewed as the sample of a stationary time series which depends on the *rescaled location* u. Note that u is a continuous time parameter, sometimes referred to as the *rescaled time* index. Following Dahlhaus (1996b), it is classical to set T as the number of available observations, in which case all the definitions are restricted to $1 \le t \le T$ and $u \in [0, 1]$. However this is not essential in the mathematical derivations and it is more convenient to set $t \in \mathbb{Z}$ and $u \in \mathbb{R}$ for presenting our setting.
We first introduce definitions for the time varying covariance and the local covariance functions.

Definition 12 (Time varying covariance function)**.** *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ be an array of random variables with finite variances. The local time varying covariance function γ^* is defined for all $t \in \mathbb{Z}, T \in \mathbb{N}^*$ and $\ell \in \mathbb{Z}$ as*

$$\gamma^*(t, T, \ell) = \operatorname{cov}(X_{t,T}, X_{t-\ell,T}) \ . \tag{5.2.1}$$

Definition 13 (Local covariance function)**.** *A* local spectral density *f is a $\mathbb{R}^2 \to \mathbb{R}_+$ function, (2π)-periodic and locally integrable with respect to the second variable. The* local covariance function *γ associated with the time varying spectral density f is defined on $\mathbb{R} \times \mathbb{Z}$ by*

$$\gamma(u, \ell) = \int_{-\pi}^{\pi} \exp(\mathrm{i}\ell\lambda)\, f(u, \lambda)\, \mathrm{d}\lambda \ . \tag{5.2.2}$$

In (5.2.2), the variable u should be seen as *rescaled* time index (in $\mathbb{R}$), ℓ as a (non-rescaled) time index and λ as a frequency (in $[-\pi, \pi]$). Recall that, by the Herglotz theorem (see (Brockwell and Davis, 2002, Theorem 4.3.1)), Equation (5.2.2) guaranties that for any $u \in \mathbb{R}$, $(\gamma(u, \ell))_{\ell\in\mathbb{Z}}$ is indeed the autocovariance function of a stationary time series. Now, we can state the definition of locally stationary processes that we use here.

Definition 14 (Locally stationary processes)**.** *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ be an array of random variables with finite variances. We say that $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ is* locally stationary with local

spectral density f *if the time varying covariance function* γ^* *of* $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ *and the local covariance function* γ *associated with* f *satisfy*

$$\left|\gamma^*(t,T,\ell) - \gamma\left(\frac{t}{T},\ell\right)\right| \leq \frac{C}{T}\ , \tag{5.2.3}$$

where C *is a constant.*

Let us give some examples fulfilling this definition.

Example 15. Locally stationary processes were first introduced by Dahlhaus (1996b) using the spectral representation

$$X_{t,T} = \int_{-\pi}^{\pi} \exp(it\omega)\, A^0_{t,T}(\omega)\, \xi(\mathrm{d}\omega)\ , \tag{5.2.4}$$

where $\xi(\mathrm{d}\omega)$ is the spectral representation of a white noise and $(A^0_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ is a collection of transfer functions such that there exist a constant K and a (unique) $2\pi-$periodic function $A : \mathbb{R}\times\mathbb{R} \to \mathbb{C}$ with $A(u,-\omega) = \overline{A(u,\omega)}$ such that for all $T \geq 1$,

$$\sup_{t,\omega}\left|A^0_{t,T}(\omega) - A\left(\frac{t}{T},\omega\right)\right| \leq \frac{K}{T}\ . \tag{5.2.5}$$

This class of locally stationary processes satisfies Definition 14 (see (Dahlhaus, 1996a, Section 1)) with $f(u,\lambda) = |A(u,\lambda)|^2$.

Example 16 (TVAR(p) model). Under suitable assumptions, the TVAR process is a particular case of Example 15 (see (Dahlhaus, 1996b, Theorem 2.3)). It is defined by the recursive equation

$$X_{t,T} = \sum_{j=1}^{p} \theta_j\left(\frac{t}{T}\right) X_{t-j,T} + \sigma\left(\frac{t}{T}\right)\xi_t\ ,$$

where $\boldsymbol{\theta} = [\theta_1\ \ldots\ \theta_p]' : \mathbb{R} \to \mathbb{R}^p$ are the time varying autoregressive coefficients and $(\xi_t)_{t\in\mathbb{Z}}$ are i.i.d. centred and with variance 1.

Example 17 (Non-stationary Causal Bernoulli Shift). Let $p > 0$ and $\varphi : \mathbb{R}^{p+2} \to \mathbb{R}$. Consider

$$X_{t,T} = \varphi\left(\frac{t}{T}, \xi_t, \ldots, \xi_{t-p}\right)\ ,$$

where $(\xi_t)_{t\in\mathbb{Z}}$ are i.i.d. such that $\mathbb{E}[|\xi_0|^q] < \infty$ for all $q \geq 1$, $\mathbb{E}[\varphi(u, \xi_0, \ldots, \xi_p)] = 0$ for all $u \in \mathbb{R}$ and there exist $K, C, r > 0$ such that, for all $u, u' \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^{p+1}$

$$\begin{aligned} |\varphi(u, \boldsymbol{x})| &\leq C\left(1 + \sum_{i=0}^{p} |x_i|^r\right), \\ \left|\varphi(u, \boldsymbol{x}) - \varphi(u', \boldsymbol{x})\right| &\leq K|u - u'|\left(1 + \sum_{i=0}^{p} |x_i|^r\right). \end{aligned}$$

In contrast to Examples 15 and 16, Example 17 do not rely on a linear representation of the process.

5.2.2 Statement of the problem

Let $d \in \mathbb{N}^*$. For each $t = 1, \ldots, T$, define the prediction vector of order d by

$$\boldsymbol{\theta}^*_{t,T} = \underset{\boldsymbol{\theta}=[\theta_1 \ldots \theta_d]' \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\left[\left(X_{t,T} - \sum_{k=1}^{d} \theta_k\, X_{t-k,T}\right)^2\right] = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}\left[\left(X_{t,T} - \boldsymbol{\theta}'\mathbf{X}_{t-1,T}\right)^2\right], \quad (5.2.6)$$

where A' denotes the transpose of matrix A and $\mathbf{X}_{s,T} = [X_{s,T} \ \ldots \ X_{s-(d-1),T}]'$. Provided that $\Gamma^*_{t,T}$ is invertible, the solution is given by

$$\boldsymbol{\theta}^*_{t,T} = \left(\Gamma^*_{t,T}\right)^{-1} \boldsymbol{\gamma}^*_{t,T}, \quad (5.2.7)$$

where $\boldsymbol{\gamma}^*_{t,T} = [\gamma^*(t, T, 1) \ \ldots \ \gamma^*(t, T, 1)]'$, $\Gamma^*_{t,T}$ is the time varying covariances matrix $\Gamma^*_{t,T} = (\gamma^*(t - i, T, j - i);\ i, j = 1, \ldots, d)$ and γ^* is the time varying covariance function as defined in (5.2.1). Analogously to (5.2.7), and with the aim of approximating the local solution of the stationary Yule-Walker equations, we set

$$\boldsymbol{\theta}_u = \Gamma_u^{-1}\boldsymbol{\gamma}_u, \quad (5.2.8)$$

where $\boldsymbol{\gamma}_u = [\gamma(u, 1) \ \ldots \ \gamma(u, d)]'$, Γ_u is the covariances matrix $\Gamma_u = (\gamma(u, i - j);\ i, j = 1, \ldots, d)$ and γ is the local covariance function as defined in (5.2.2).
Assuming particular regularity conditions on $\boldsymbol{\theta}$, an estimator $\widehat{\boldsymbol{\theta}}$ of it is studied in Dahlhaus and Giraitis (1998) for the model of Example 15. In the following we improve these results by deriving minimax rate properties of the estimator of Dahlhaus and Giraitis (1998) and extensions of it. We will use the following smoothness class of functions. For $\alpha \in (0, 1]$ the $\alpha-$Hölder semi-norm of a function $\mathbf{f} : \mathbb{R} \to \mathbb{C}^d$ is defined by

$$|\mathbf{f}|_{\alpha,0} = \sup_{0<|s-s'|<1} \frac{\|\mathbf{f}(s) - \mathbf{f}(s')\|}{|s - s'|^\alpha}.$$

This semi-norm is used to build a norm for any $\beta > 0$ as it follows. Let $k \in \mathbb{N}$ and $\alpha \in (0, 1]$ be such that $\beta = k + \alpha$. If $\mathbf{f}$ is k times differentiable on $\mathbb{R}$, we define

$$|\mathbf{f}|_\beta = \left|\mathbf{f}^{(k)}\right|_{\alpha,0} + \max_{0\le s\le k} \left\|\mathbf{f}^{(s)}\right\|_\infty ,$$

and $|\mathbf{f}|_\beta = \infty$ otherwise. For $R > 0$ and $\beta > 0$, the $(\beta, R)-$ Hölder ball of dimension d is denoted by

$$\Lambda_d(\beta, R) = \left\{\mathbf{f} : \mathbb{R} \to \mathbb{C}^d, \text{ such that } |\mathbf{f}|_\beta \le R\right\} .$$

We can now derive the main assumption used on the model which depends on some positive constants β, R and f_-.

(**M**-3) The sequence $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ is a locally stationary process in the sense of Definition 14 such that $\mathbb{P}(X_{t,T} = 0) = 0$ for any t. The spectral density $f(\cdot, \lambda)$ belongs to $\Lambda_1(\beta, R)$ for any $\lambda \in \mathbb{R}$, and satisfies $f(u, \lambda) \ge f_-$ for all $u, \lambda \in \mathbb{R}$. The constant C in (5.2.3) depends continuously and at most on $\|f\|_\infty$ and $\sup_{u,\lambda\in\mathbb{R}} |\partial f(u, \lambda)/\partial u|$.

Note in particular that for $\beta > 1$, (**M**-3) implies that f is continuously differentiable in its first component.

The problem that we are interested in is to derive a minimax rate estimator $\widetilde{\boldsymbol{\theta}}$ for any $\beta \ge 2$, which means, that for such a β, the estimation risk, say the quadratic risk $\mathbb{E}[\|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}^*_{t,T}\|^2]$ can be bounded uniformly over all processes satisfying (**M**-3) (among with some additional assumptions), and that the corresponding rate of convergence as $T \to \infty$ cannot be improved by any other estimator. The case $\beta \le 2$ is solved in Moulines et al. (2005) for a particular subclass.

5.3 MINIMAX ESTIMATION FOR ADAPTIVE PREDICTION

Let $\widehat{X}^*_{d,t,T}$ denote the best linear predictor of order d of $X_{t,T}$, which as a consequence of (5.2.6), reads

$$\widehat{X}^*_{d,t,T} = \left(\boldsymbol{\theta}^*_{t,T}\right)' \mathbf{X}_{t-1,T} ,$$

We denote by $\widehat{X}^*_{t,T}$ the best predictor of $X_{t,T}$ given its past, which corresponds to the conditional expectation

$$\widehat{X}^*_{t,T} = \mathbb{E}\left[X_{t,T} \left| X_{s,T},\ s \le t-1\right.\right] . \tag{5.3.1}$$

As explained before, the goal of this paper is to derive estimators, say $\widetilde{\boldsymbol{\theta}}_{t,T} \in \mathbb{R}^d$, of $\boldsymbol{\theta}_{t/T}$, which is a local approximation of $\boldsymbol{\theta}^*_{t,T}$. In this section, we assume that $\widetilde{\boldsymbol{\theta}}_{t,T}$ is a function of the past $X_{s,T}$, $s \le t-1$. Then $\widetilde{\boldsymbol{\theta}}'_{t,T}\mathbf{X}_{t-1,T}$ is a legitimate predictor of $X_{t,T}$ and we have the following decomposition of the corresponding prediction quadratic risk

$$\mathbb{E}\left[\left(X_{t,T} - \widetilde{\boldsymbol{\theta}}'_{t,T}\mathbf{X}_{t-1,T}\right)^2\right] = \mathbb{E}\left[\left(X_{t,T} - \widehat{X}^*_{t,T}\right)^2\right] + \mathbb{E}\left[\left(\widetilde{\boldsymbol{\theta}}'_{t,T}\mathbf{X}_{t-1,T} - \widehat{X}^*_{t,T}\right)^2\right] .$$

The first term is the minimal prediction error that one would achieve with the conditional expectation (which requires the true distribution of the whole process). Furthermore, inserting $\widehat{X}^*_{d,t,T} = \left(\boldsymbol{\theta}^*_{t,T}\right)' \mathbf{X}_{t-1,T}$ and using the Minkowskii and Cauchy-Schwartz inequality, the square root of the second term can be bounded as

$$\left(\mathbb{E}\left[\left(\widetilde{\boldsymbol{\theta}}'_{t,T}\mathbf{X}_{t-1,T} - \widehat{X}^*_{t,T}\right)^2\right]\right)^{1/2} \leq \left(\mathbb{E}\left[\left(\widehat{X}^*_{d,t,T} - \widehat{X}^*_{t,T}\right)^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\left\|\mathbf{X}_{t-1,T}\right\|^4\right]\right)^{1/4} \left(\mathbb{E}\left[\left\|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}^*_{t,T}\right\|^4\right]\right)^{1/4} .$$

The first term in the upper bound is due to the approximation of the best predictor by the best linear predictor of order d and can only be improved by increasing d. Note that, in the case of the TVAR(p) model with $p \leq d$, this error term vanishes. The quantity $(\mathbb{E}[\|\mathbf{X}_{t-1,T}\|^2])^{1/2}$ is typically bounded by a universal constant independent of (t, T) over the class of processes under consideration. Hence, for a given d, the control of the prediction risk boils down to the control of the quadratic estimation risk $\mathbb{E}[\|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}^*_{t,T}\|^2]$.
To do so, we can further decompose the quadratic loss as

$$\left\|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}^*_{t,T}\right\| \leq \left\|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}_{t/T}\right\| + \left\|\boldsymbol{\theta}_{t/T} - \boldsymbol{\theta}^*_{t,T}\right\| ,$$

and note that the second term is a deterministic error basically accounting for the approximation precision of the non-stationary model by a stationary one, which, under appropriate assumptions, will be shown to be of order T^{-1}. As a result of these successive decompositions, our effort in the following focus on controlling the estimation risk $\mathbb{E}[\|\widetilde{\boldsymbol{\theta}}_{t,T} - \boldsymbol{\theta}_{t/T}\|^2]$ uniformly over a class of locally stationary processes with given smoothness index $\beta \geq 2$.
By achieving this goal, we will provide a theoretical justification of the intuitive fact that, in a non-stationary context, any predictor should be adapted to how smoothly the time varying parameter evolves along the time. On the other hand, in practical situations, one may not have a strong *a priori* on the smoothness index β and one should rely on data driven methods that are therefore called *adaptive*. This problem was tackled in Chapter 4. More precisely, using aggregation techniques introduced in the context of individual sequences prediction (see Vovk (1990); Littlestone and Warmuth (1994); Cesa-Bianchi and Lugosi (2006); Anava et al. (2013)) and statistical learning (Barron (1987); Catoni (1997, 2004); Juditsky and Nemirovski (2000); Yang (2000a, 2004); Leung and Barron (2006)), one can aggregate sufficiently many predictors in order to build a minimax predictor which adapts to the unknown smoothness β of the time varying parameter. However, a crucial requirement in Chapter 4 is to dispose of β-minimax-rate sequences of predictors for any $\beta > 0$. Hence, following Chapter 4 and Moulines et al. (2005), where minimax estimators are derived only for $\beta \leq 2$, our results will pave the way for adaptive minimax-rate forecasting at any (unknown) smoothness rate.

5.4 TAPERED YULE-WALKER ESTIMATE

Following Dahlhaus and Giraitis (1998), a local empirical covariance function is defined as follows. It relies on a real data taper function h and a bandwidth M which may depend on T.

Definition 15 (Empirical local covariance function)**.** *Consider a function $h : [0,1] \to \mathbb{R}$ and $M \in 2\mathbb{N}^*$. The empirical local covariance function $\widehat{\gamma}_M$ with taper h is defined in $\mathbb{R}\times\mathbb{Z}$ as*

$$\widehat{\gamma}_M(u,\ell) = \frac{1}{H_M} \sum_{\substack{t_1,t_2=1 \\ t_1-t_2=\ell}}^{M} h\left(\frac{t_1}{M}\right) h\left(\frac{t_2}{M}\right) X_{\lfloor uT\rfloor+t_1-M/2,T} X_{\lfloor uT\rfloor+t_2-M/2,T} \ ,$$

where $H_M = \sum_{k=1}^{M} h^2(k/M) \sim M \int_0^1 h^2(x)\mathrm{d}x$ is the normalizing factor. We assume that $H_M > 0$.

5

For $h \equiv 1$ in Definition 15 we obtain the classical covariance estimate for a centred sample $\{X_s, \lfloor uT\rfloor - M/2 \le s \le \lfloor uT\rfloor + \ell + M/2\}$. Taking into account the interval $[t - M/2 + 1, t + M/2]$, and with the help of the data taper function h, the following empirical Yule-Walker equations are then derived

$$\widehat{\boldsymbol{\theta}}_{t,T}(M) \ = \ \widehat{\Gamma}^{-1}_{t,T,M} \widehat{\boldsymbol{\gamma}}_{t,T,M} \ , \tag{5.4.1}$$

where $\widehat{\boldsymbol{\gamma}}_{t,T,M} = [\widehat{\gamma}_M(t/T,1) \ \ldots \ \widehat{\gamma}_M(t/T,d)]'$, $\widehat{\Gamma}_{t,T,M}$ is the matrix of empirical covariances $\widehat{\Gamma}_{t,T,M} = (\widehat{\gamma}_M(t/T, i-j);\ i,j = 1,\ldots,k)$ and $\widehat{\gamma}_M$ is the empirical covariance function as in Definition 15.

5.5 MAIN RESULTS IN THE GENERAL FRAMEWORK

5.5.1 Additional assumptions

For convenience, we introduce the following notation. Let $p > 0$, $q, r, s \in \mathbb{N}^*$, $u : \mathbb{R} \to \mathbb{R}$, $a, b : \mathbb{R}^r \to \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^q$ and a collection of random matrices $\{U_M \in \mathbb{R}^{r\times s}, M \in \mathbb{N}^*\}$. We write

(i) $U_M = O_{L^p,\boldsymbol{c}}(u(M))$ if there exists $C_{p,\boldsymbol{c}} > 0$, depending continuously and at most on $(p, \boldsymbol{c}')$, such that for all $M \in \mathbb{N}^*$

$$\max_{1\le i\le r, 1\le j\le s} \left(\mathbb{E}\left[\left|U_{M,i,j}\right|^p\right]\right)^{1/p} \le C_{p,\boldsymbol{c}} \left|u(M)\right| \ , \tag{5.5.1}$$

where $U_{M,i,j}$ is the (i,j)-th entry of the matrix U_M.

(ii) $U_M = O_{L^\bullet,\boldsymbol{c}}(u(M))$ if $U_M = O_{L^p,\boldsymbol{c}}(u(M))$ for all $p > 0$.

(iii) $a(\boldsymbol{x}) = O_{\boldsymbol{c}}(b(\boldsymbol{x}))$ if and only if there exists a constant $C_{\boldsymbol{c}}$ depending continuously and at most on the index $\boldsymbol{c}$, such that for all $\boldsymbol{x} \in \mathbb{R}^r$

$$|a(\boldsymbol{x})| \leq C_{\boldsymbol{c}} |b(\boldsymbol{x})| \ .$$

Concerning the function h we have the following assumption.

(**H**) The function $h : [0,1] \to \mathbb{R}$ is piecewise continuously differentiable, that is, for $0 = u_0 < u_1 < \ldots < u_N = 1$, h is C^1 on $(u_{i-1}, u_i]$, $i = 1, \ldots, N$. Moreover we denote $\|h\|_\infty = \sup_{u\in[0,1]} |h(u)|$ and $\|h'\|_\infty = \max_{1\leq i\leq N} \sup_{u\in(u_{i-1},u_i]} |h'(u)|$.

Provided a piecewise continuously differentiable funtion h (as in (**H**)) and a local spectral density function f continuously differentiable on its first component, we also consider the following assumption.

(**C**) For all $q > 0$, $M_q := \sup_{t,T} \mathbb{E}\left[\left|X_{t,T}\right|^q\right] < \infty$ and for all $\ell \in \mathbb{Z}$ the empirical covariance function satisfy

$$\gamma_M(u,\ell) - \mathbb{E}\left[\gamma_M(u,\ell)\right] = O_{L^\bullet,\ell,f_-,\|h\|_\infty,\|h'\|_\infty,\|f\|_\infty,\|\partial f/\partial u\|_\infty}\left(M^{-1/2}\right) \ .$$

At first glance Assumption (**C**) may seem restrictive but it is not. Locally stationary processes of Example 15 satisfy it (see (Dahlhaus and Giraitis, 1998, Theorem 4.1)) and also m-dependent sequences as those in Example 17.

5.5.2 Bound of the estimation risk

Our first result provides an equality satisfied by the estimation error of $\widehat{\boldsymbol{\theta}}_{t,T}(M)$.

Theorem 5.5.1. *Let $\beta \geq 2, R, f_- > 0$ and $h : [0,1] \to \mathbb{R}$. Let $k \in \mathbb{N}$ and $\alpha \in (0,1]$ be uniquely defined such that $\beta = k + \alpha$ and consider $M \in 2\mathbb{N}^*$. Suppose that Assumptions (**M**-3), (**H**) and (**C**) hold. Let $\widehat{\boldsymbol{\theta}}_{t,T}(M)$ be obtained from Equation (5.4.1). The following relation is satisfied*

$$\widehat{\boldsymbol{\theta}}_{t,T}(M) - \boldsymbol{\theta}_{t/T} = \sum_{\ell=1}^{k} \boldsymbol{a}_{h,f,\ell} \left(\frac{M}{T}\right)^\ell + O_{d,f_-,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^\beta\right) + \boldsymbol{v}_M \ , \tag{5.5.2}$$

where $\boldsymbol{a}_{h,f,\ell}$ depends only on h, the spectral density f and ℓ and $\boldsymbol{v}_M = O_{L^\bullet,d,f_-,\|h\|_\infty,\|h'\|_\infty,\beta,R}(M^{-1/2})$. Moreover, $\boldsymbol{a}_{h,f,1} = 0$ if $h(x) = h(1-x)$ for $x \in [0,1]$.

The proof can be found in Section 5.9.1. Theorem 5.5.1 suggests to combine several $\widehat{\boldsymbol{\theta}}_{t,T}(M)$ to obtain a more accurate estimation by cancelling out the first k bias terms in (5.5.2). The technique was already used for eliminate one term of bias in (Moulines et al., 2005, Theorem 8) for example. It is inspired by the Romberg's method in numerical analysis (see Baranger and Brezinski (1991)). Let $\boldsymbol{\alpha} = [\alpha_0 \ \ldots \ \alpha_k]' \in \mathbb{R}^{k+1}$, be the solution of the equation

$$A\boldsymbol{\alpha} = \boldsymbol{e}_1 \ , \tag{5.5.3}$$

where $\boldsymbol{e}_1 = [1 \ 0 \ \ldots \ 0]'$ is the $\mathbb{R}^{k+1}$- vector having a 1 in the first position and zero everywhere else and A is a $(k+1)\times(k+1)$ matrix with entries $A_{i,j} = 2^{-ij}$ for $0 \leq i, j \leq k$.

Theorem 5.5.2. *Let $\beta \geq 2, R, f_- > 0$ and $h : [0,1] \to \mathbb{R}$. Let $k \in \mathbb{N}$ and $\alpha \in (0,1]$ be uniquely defined such that $\beta = k + \alpha$ and consider $M \in 2^{k+1}\mathbb{N}^*$. Suppose that Assumptions (M-3), (H) and (C) hold. Let $\widehat{\boldsymbol{\theta}}_{t,T}(M)$ be obtained from Equation (5.4.1). Then, $\widetilde{\boldsymbol{\theta}}_{t,T}(M) = \sum_{\ell=0}^{k} \alpha_\ell \widehat{\boldsymbol{\theta}}_{t,T}(M/2^\ell)$ with $\boldsymbol{\alpha}$ defined by (5.5.3) satisfies*

$$\widetilde{\boldsymbol{\theta}}_{t,T}(M) - \boldsymbol{\theta}_{t/T} = O_{d,f_-,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^\beta\right) + \boldsymbol{v}_M \,, \tag{5.5.4}$$

where $\boldsymbol{v}_M = O_{L^\bullet,d,f_-,\|h\|_\infty,\|h'\|_\infty,\beta,R}(M^{-1/2})$.

The proof is postponed to Section 5.9.2. It is straightforward to check that the optimal bandwidth for minimizing the order of the right term of Equation (5.5.4) is $M \propto T^{2\beta/(2\beta+1)}$. The next result is a direct consequence of Lemma 17, Theorem 5.5.2 and this observation.

5

Corollary 2. *Let $\beta \geq 2, R, f_- > 0$ and $h : [0,1] \to \mathbb{R}$. Let $k \in \mathbb{N}$ and $\alpha \in (0,1]$ be uniquely defined such that $\beta = k + \alpha$ and consider $M = 2^{k+1}\lfloor T^{2\beta/(2\beta+1)} \rfloor$. Suppose that Assumptions (M-3), (H) and (C) hold. Let $\widetilde{\boldsymbol{\theta}}_{t,T}(M)$ be obtained as in Theorem 5.5.2. Then, for any $q > 0$ there exist a constant C only depending on h, q, d, f_-, R and continuously on β and a $T_0 > 0$ only depending on d, R and f_- such that, if $T \geq T_0$ we have, for all $t \in \mathbb{Z}$,*

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{\theta}}_{t,T}(M) - \boldsymbol{\theta}^*_{t,T}\right\|^q\right] \leq \frac{C}{T^{q\beta/(2\beta+1)}} \,. \tag{5.5.5}$$

5.6 Application to TVAR processes

Time varying autoregressive processes (see Example 16) are a handful model to illustrate our results.

The index β sets the regularity of the functions we are interested in (the TVAR parameter $\boldsymbol{\theta}$). The following concepts are related to standard stability conditions on them.

For $\boldsymbol{\theta} : \mathbb{R} \to \mathbb{R}^p$, we define the time varying autoregressive polynomial by $\boldsymbol{\theta}(z;u) = 1 - \sum_{j=1}^{p} \theta_j(u) z^j$.

Let us denote, for any $\delta > 0$, $s_p(\delta) = \{\boldsymbol{\theta} : \mathbb{R} \to \mathbb{R}^p, \boldsymbol{\theta}(z;u) \neq 0, \forall |z| < \delta^{-1}, u \in [0,1]\}$.

Define, for $\beta > 0$, $R > 0$, $\delta \in (0,1)$, $\rho \in [0,1]$ and $\sigma_+ > 0$, the class of parameters

$$C(\beta, R, \delta, \rho, \sigma_+) = \left\{(\boldsymbol{\theta}, \sigma) : \mathbb{R} \to \mathbb{R}^p \times [\rho\sigma_+, \sigma_+] \ : \ \boldsymbol{\theta} \in \Lambda_p(\beta, R) \cap s_p(\delta)\right\} \,.$$

Given an i.i.d. sequence $(\xi_t)_{t\in\mathbb{Z}}$ and constants $\delta \in (0,1)$, $\rho \in [0,1]$, $\sigma_+ > 0$, $\beta > 0$ and $R > 0$, we consider the following assumption.

(**M**-4) The sequence $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ is a TVAR process with time varying standard deviation σ, time varying AR coefficients $\theta_1, \ldots, \theta_p$ and innovations $(\xi_t)_{t\in\mathbb{Z}}$, and $(\boldsymbol{\theta}, \sigma) \in C(\beta, R, \delta, \rho, \sigma_+)$.

A TVAR process admits a linear representation with respect to the innovations (see Proposition 2 in Chapter 4). It is convenient to introduce the assumption below.

(**I**) For all $q > 0$ the innovations $(\xi_t)_{t\in\mathbb{Z}}$ satisfy $m_q := \mathbb{E}\left[|\xi|^q\right] < \infty$.

Time varying autoregressive processes are locally stationary under certain conditions on their parameters and moments. The next result is consequence of (Dahlhaus, 1996b, Theorem 2.3).

Theorem 5.6.1. *Let $\delta \in (0,1), \beta > 0, R > 0$ and $\rho \in [0,1]$. Suppose that Assumptions (**M**-4) and (**I**) hold. Then, the process is locally stationary in the sense of Definition 14 with*

$$f(u,\lambda) = \frac{\sigma^2(u)}{2\pi}\left(1 - \sum_{j=1}^{p} \theta_j(u)\exp(-ij\lambda)\right)^{-2} . \tag{5.6.1}$$

Moreover, $\boldsymbol{\theta}^*_{t,T} \in \mathbb{R}^d$ as defined by Equation (5.2.8) coincides with $\boldsymbol{\theta}(t/T)$ when $p = d$. To apply the results of Section 5.5 to the TVAR fulfilling Assumption (**M**-4) and (**I**), we should take care of the regularity of the spectral density and also of its bounds. The analysis of Section 5.10 points in that direction. From that, we conclude that the conditions of Corollary 2 are fulfilled.

Corollary 3. *Let $\delta \in (0,1), \beta \geq 2, R > 0$ and $\rho \in [0,1]$. Let $k \in \mathbb{Z}$ and $\alpha \in (0,1]$ be uniquely defined such that $\beta = k + \alpha$ and consider $M = 2^{k+1}\lfloor T^{2\beta/(2\beta+1)}\rfloor$. Suppose that Assumptions (**M**-4) and (**I**) hold and that $\mathbb{P}(X_{t,T} = 0) = 0$ for any t. Assume moreover that $\sigma \in \Lambda_1(\beta, R)$. Let $\widetilde{\boldsymbol{\theta}}_{t,T}(M)$ be a p dimensional vector obtained as in Theorem 5.5.2 (i.e. $p = d$). Then, for any $q \in \mathbb{N}$ there exists a constant C only depending on $q, h, p, \delta, \rho, \sigma_+, R$ and continuously on β, and $T_0 > 0$ depending only on $p, \delta, \rho, \sigma_+, R$ and β such that, for $T \geq T_0$ we have*

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{\theta}}_{t,T}(M) - \boldsymbol{\theta}\left(\frac{t}{T}\right)\right\|^q\right] \leq \frac{C}{T^{q\beta/(2\beta+1)}} . \tag{5.6.2}$$

The estimator $\widetilde{\boldsymbol{\theta}}$ proposed in Corollary 3 is β-minimax-rate for TVAR processes according to (Moulines et al., 2005, Theorem 4). Hence, it is also β-minimax-rate in the class of locally stationary processes satisfying Assumption (**M**-3). Section 4.7.1 of Chapter 4 explains how to construct minimax-rate predictors from minimax-rate estimators of $\boldsymbol{\theta}$. Applying their approach, Corollary 3 also provides a crucial ingredient in building β-minimax-rate predictors for any $\beta \geq 2$.

5.7 NUMERICAL WORK

We test both methods on data simulated according to a TVAR process with $p = 3$. The smooth parameter function $t \mapsto \boldsymbol{\theta}(t)$ within $s_p(\delta)$ for some $\delta \in (0,1)$ is chosen as follows. First we pick randomly some smoothly time varying partial autocorrelation functions up to the order p that are bounded between -1 and 1, $\theta_{k,k}(u) = \delta^k \sum_{j=1}^{F-1} a_{j,k} j^2 \cos(ju)/[F(F-1)(2F-1)/6]$, where $a_{j,k}$ are random numbers

in $[-1, 1]$, the same ones for all u. Then we use Algorithm 7 and set $\boldsymbol{\theta} = -[\theta_{1,p} \ \dots \ \theta_{p,p}]$. From the classical Levinson-Durbin recurrence (i.e. Algorithm 7 with $\delta = 1$) we obtain a function in $s_p(1)$ (see for example Makhoul (1975)), it is straightforward to check that the $\boldsymbol{\theta}$ produced by Algorithm 7 with $\delta \in (0, 1)$ is in $s_p(\delta)$. The three components of our $\boldsymbol{\theta}(t)$ are displayed in Figure 5.1. The generated $\boldsymbol{\theta}$ is, in theory, C^∞. We can then ensure that

Algorithm 7: Adapted Levinson-Durbin algorithm.

parameters the stability parameter $\delta > 0$ and the time varying partial autocorrelation functions $\theta_{k,k}$, $k = 1, \dots, p$;

for $k = 2$ **to** p **do**

  for $j = 2$ **to** $p - 1$ **do**

    $\theta_{j,k} = \theta_{j,k-1} + \delta^{2j-2k}\theta_{k,k}\theta_{k-j,k-1}$;

for any $\beta > 0$, it is in $\Lambda_p(\beta, R)$ for some $R > 0$. For convenience we build $\widetilde{\boldsymbol{\theta}}$ with $k = 1$. For each $T \in \{2^{2j}, j = 5, \dots, 15\}$ we generate 100 realizations of a TVAR process from innovation sequences $(\xi_t)_{t\in\mathbb{Z}}$ of i.i.d. centred Gaussian random variables with unit variance by sampling $\boldsymbol{\theta}$ at a rate T^{-1}. Then we compare $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ for estimating $\boldsymbol{\theta}(1/2)$ using $h \equiv 1$ and different values of M. Recall that $\boldsymbol{\theta}(1/2) = \boldsymbol{\theta}^*_{T/2,T}$. Figure 5.2 shows the boxplots corresponding to this evaluation for two different Ts.

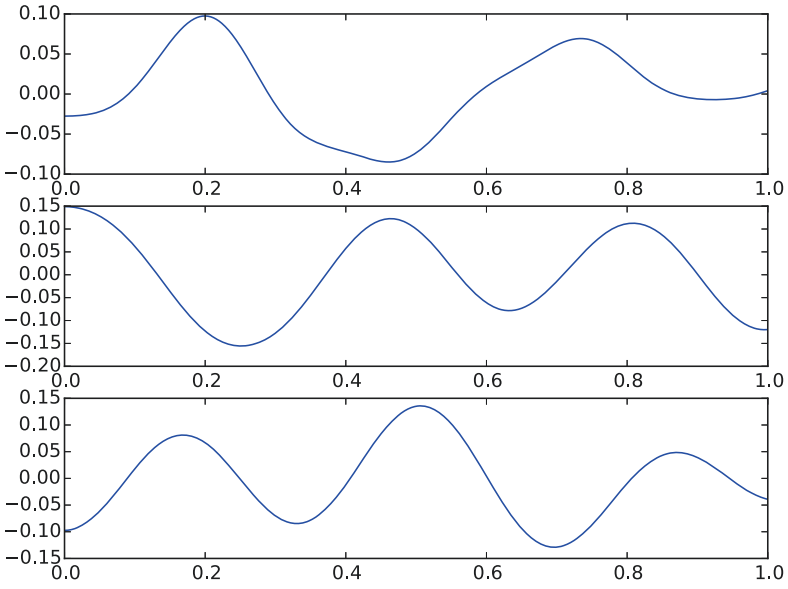


Figure 5.1 : Plots of $\theta_1(t)$ (top), $\theta_2(t)$ (middle) and $\theta_3(t)$ (bottom) on the interval $t \in [0, 1]$.

In Figure 5.2 we observe that for $T = 2^{20}$ the error of $\widehat{\boldsymbol{\theta}}$ is minimized in $M = 2^{15}$ while that of $\widetilde{\boldsymbol{\theta}}$ reaches its minimum in $M = 2^{17}$. The estimator $\widetilde{\boldsymbol{\theta}}$ beats $\widehat{\boldsymbol{\theta}}$ for the two biggest

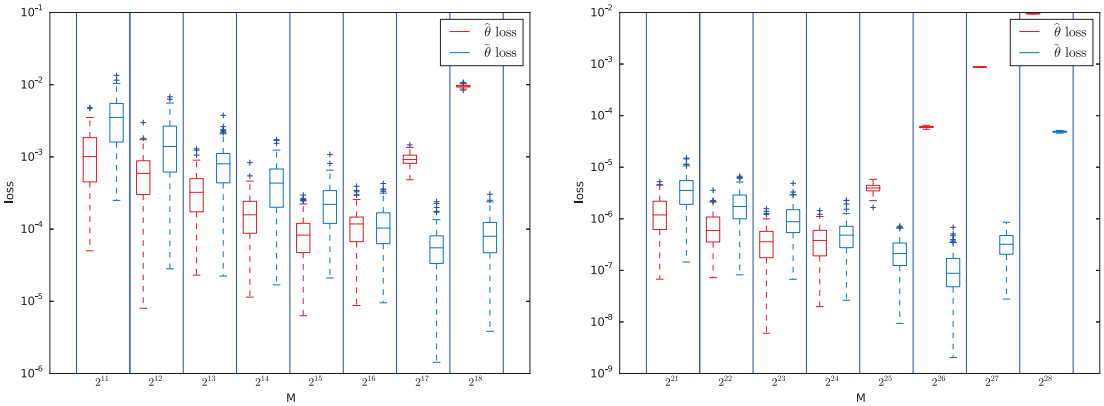


Figure 5.2 : Box plots of the quadratic losses for estimating $\boldsymbol{\theta}(1/2)$ using $\widehat{\boldsymbol{\theta}}_{T/2,T}(M)$ (red boxes) and $\widetilde{\boldsymbol{\theta}}_{T/2,T}(M)$ (blue boxes) for various bandwidths M, when $T = 2^{20}$ (left) and $T = 2^{30}$ (right).

values of M. In the case $T = 2^{30}$, the error of $\widehat{\boldsymbol{\theta}}$ reaches its minimum in $M = 2^{23} = T^{4/5}/2$ and that of $\widetilde{\boldsymbol{\theta}}$ in $M = 2^{26} = 2^2\, T^{4/5}$. The estimator $\widetilde{\boldsymbol{\theta}}$ beats $\widehat{\boldsymbol{\theta}}$ for the four biggest values of M. These experiences illustrate the theoretical result established in (Dahlhaus and Giraitis, 1998, Theorem 2.2) (where an optimal rate for $\widehat{\boldsymbol{\theta}}$ estimation is obtained with $M \propto T^{4/5}$) and Corollary 3 (exhibiting the optimal rate for $\widetilde{\boldsymbol{\theta}}$ estimation in $M \propto T^{4/5}$, if $\beta = 2$). Figure 5.3 (left graph) displays the oracle errors $\min_M \|\widehat{\boldsymbol{\theta}}_{T/2,T}(M) - \boldsymbol{\theta}(1/2)\|$ and

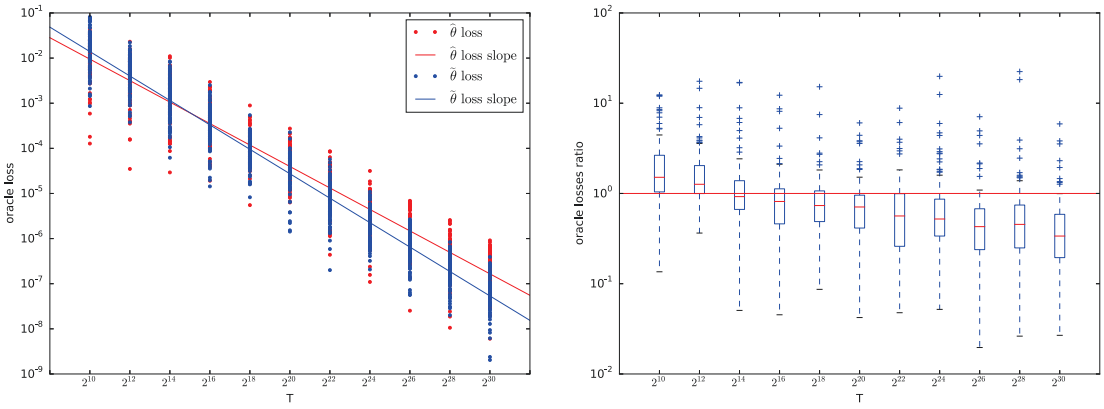


Figure 5.3 : Oracle losses (using the best choice for the bandwidth M) for estimating $\boldsymbol{\theta}(1/2)$ using $\widehat{\boldsymbol{\theta}}_{T/2,T}(M)$ (red points) and $\widetilde{\boldsymbol{\theta}}_{T/2,T}(M)$ (blue points) for various values of T. The left plot displays the losses over all the Monte Carlo simulations and the two resulting log-log regression lines. The right plot displays boxplots of the corresponding losses ratio.

$\min_M \|\widetilde{\boldsymbol{\theta}}_{T/2,T}(M) - \boldsymbol{\theta}(1/2)\|$ for all $T \in \{2^{2j}, j = 5, \ldots, 15\}$. The slope corresponding to $\widetilde{\boldsymbol{\theta}}$ (in blue) is steeper than the one corresponding to $\widehat{\boldsymbol{\theta}}$ (in red), meaning that, in average, $\widetilde{\boldsymbol{\theta}}$ outperforms $\widehat{\boldsymbol{\theta}}$ by an increasing order of magnitude as T increases. This corroborates what is expected from our theoretical analysis (see Corollary 3). The boxplots of Figure 5.3

(right graph) represent the ratios $\min_M \|\widetilde{\boldsymbol{\theta}}_{T/2,T}(M) - \boldsymbol{\theta}(1/2)\| / \min_M \|\widehat{\boldsymbol{\theta}}_{T/2,T}(M) - \boldsymbol{\theta}(1/2)\|$ computed for each T and realization of the TVAR process. Observe that for $2^{14} \leq T \leq 2^{18}$ the estimator $\widetilde{\boldsymbol{\theta}}$ beats $\widehat{\boldsymbol{\theta}}$ in at least half of the cases. For $T \geq 2^{20}$, it happens in at least 75% of the cases.

5.8 USEFUL RESULTS ON LOCALLY STATIONARY TIME SERIES

This section provides the background necessary to support the proof of our main results about locally stationary processes. The next two lemmas allow to control the norms of $\widehat{\boldsymbol{\theta}}_{t,T}$ and $\boldsymbol{\theta}_{t/T}$.

Lemma 15. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ be a locally stationary process in the sense of Definition 14 such that $\mathbb{P}(X_{t,T} = 0) = 0$ for any $t \leq T$. The Yule-Walker estimate $\widehat{\boldsymbol{\theta}}_{t,T}(M)$ defined by Equation (5.4.1) satisfies $\|\widehat{\boldsymbol{\theta}}_{t,T}(M)\| \leq 2^d$ almost surely.*

Proof. This proof is an adaptation of that of (Dahlhaus and Giraitis, 1998, Lemma 4.2). We start by showing that $\widehat{\Gamma}_{t,T,M}$, with entries defined as in (5.4.1), is non-singular almost surely. Suppose on the contrary that $\mathbb{P}(\det(\widehat{\Gamma}_{t,T,M}) = 0) > 0$. This means that there is an $\boldsymbol{x} \in (\mathbb{R}^d)^*$ such that $\widehat{\Gamma}_{t,T,M}\boldsymbol{x} = 0$ and therefore $\boldsymbol{x}'\widehat{\Gamma}_{t,T,M}\boldsymbol{x} = \int_{-\pi}^{\pi} \widehat{f}_M(t/T,\lambda)|\sum_{j=1}^d x_j \exp(ij\lambda)|^2 \mathrm{d}\lambda = 0$. The expression inside the modulus vanishes at most for $d-1$ values of λ, otherwise $\boldsymbol{x} = 0$ because the obtained Vandermonde determinant is non-zero. Then $\widehat{f}_M(t/T,\lambda) = 0$ for almost all $\lambda \in [-\pi,\pi]$. Since $\{\exp(-i\lambda s), s = 0,\ldots,M-1\}$ is a subset of an orthogonal basis of $L^2([0,1])$ we get that $h(s/M)X_{t-M/2+s+1,T} = 0$ for $s = 0,\ldots,M-1$, but then $\mathbb{P}(X_{t,T} = 0) > 0$ for some $1 \leq t \leq T$. Observe that for any s, $\widehat{\gamma}_M(s,\cdot)$ defined by (5.4.1) is an autocovariance function. Setting $s = t/T$, the corresponding covariance matrix $\widehat{\Gamma}_{t,T,M}$ is positive-definite almost surely. As consequence of Lemma 23 (Section 5.11) we have that $z_1,\ldots,z_d$, the roots of the polynomial $\widehat{\boldsymbol{\theta}}_{t,T}(z) = 1 - \sum_{j=1}^d \widehat{\theta}_{j,t,T} z^j$ satisfy $|z_j| > 1$ for any j. Then,

$$\left\|\widehat{\boldsymbol{\theta}}_{t,T}(M)\right\|^2 + 1 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|1 - \sum_{j=1}^{d}\widehat{\theta}_{j,t,T}\exp(ij\lambda)\right|^2 \mathrm{d}\lambda = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\widehat{\boldsymbol{\theta}}_{t,T}(\exp(i\lambda))\right|^2 \mathrm{d}\lambda \quad (5.8.1)$$

Note that $\prod_{j=1}^d(-z_i) = 1$. Therefore

$$\widehat{\boldsymbol{\theta}}_{t,T}(z) = \prod_{j=1}^{d}\left(z - z_j\right) = \prod_{j=1}^{d}\left(1 - zz_j^{-1}\right). \quad (5.8.2)$$

If $|z| = 1$, Equation (5.8.2) implies that $|\widehat{\boldsymbol{\theta}}_{t,T}(z)| \leq 2^d$. Putting this into (5.8.1) the proof is completed. □

Lemma 16. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ be a locally stationary process in the sense of Definition 14. Assume that $f(u,\lambda) > 0$ for all $u,\lambda \in \mathbb{R}$. The vector $\boldsymbol{\theta}_u$ defined by Equation (5.2.8) satisfies $\|\boldsymbol{\theta}_u\| \leq 2^d$.*

Proof. The proof follows the same scheme of that of Lemma 15 up to simplifications. Here the contradiction $f(u,\lambda)=0$ for almost all $\lambda\in[-\pi,\pi]$ raises immediately from the assumptions. Observe that, instead of an almost sure result, this is a deterministic one. □

Lemma 16 is necessary to prove the following.

Lemma 17. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ be a locally stationary process in the sense of Definition 14 where the spectral density f satisfies $f(u,\lambda)\geq f_-$ for all $u,\lambda\in\mathbb{R}$. Then, there exist two constants $C_1,T_0>0$ depending only on d,C (see Inequality (5.2.3)) and f_- such that, for $T\geq T_0$ we have*

$$\left\|\boldsymbol{\theta}^*_{t,T}-\boldsymbol{\theta}_{t/T}\right\|\leq\frac{C_1}{T}\,, \tag{5.8.3}$$

Proof. From equations (5.2.7) and (5.2.8) we obtain that

$$\boldsymbol{\theta}^*_{t,T}-\boldsymbol{\theta}_{t/T} \;=\; \left(\Gamma^*_{t,T}\right)^{-1}\left[\left(\Gamma_{t/T}-\Gamma^*_{t,T}\right)\boldsymbol{\theta}_{t/T}+\boldsymbol{\gamma}^*_{t,T}-\boldsymbol{\gamma}_{t/T}\right]\,.$$

Applying matrix inequalities (specifically with the spectral norm) we get

$$\|\boldsymbol{\theta}^*_{t,T}-\boldsymbol{\theta}_{t/T}\|\leq\left\|\left(\Gamma^*_{t,T}\right)^{-1}\right\|\left(\left\|\Gamma_{t/T}-\Gamma^*_{t,T}\right\|\left\|\boldsymbol{\theta}_{t/T}\right\|+\left\|\boldsymbol{\gamma}^*_{t,T}-\boldsymbol{\gamma}_{t/T}\right\|\right)\,.$$

Inequality (5.2.3) implies that $\|\Gamma_{t/T}-\Gamma^*_{t,T}\|\leq d^{3/2}C/T$ and that $\|\boldsymbol{\gamma}^*_{t,T}-\boldsymbol{\gamma}_{t/T}\|\leq d^{1/2}C/T$. The smallest eigenvalue of the matrix $\Gamma_{t/T}$ is positive, at least $2\pi f_-$ (see (Brockwell and Davis, 2006, Proposition 4.5.3)). Observe that

$$\begin{aligned}\inf_t\inf_{\|\boldsymbol{a}\|=1}\boldsymbol{a}'\Gamma^*_{t,T}\boldsymbol{a}&=\inf_t\inf_{\|\boldsymbol{a}\|=1}\left\{\boldsymbol{a}'\left(\Gamma^*_{t,T}-\Gamma_{t/T}\right)\boldsymbol{a}+\boldsymbol{a}'\Gamma_{t/T}\boldsymbol{a}\right\}\\&\geq\inf_t\inf_{\|\boldsymbol{a}\|=1}\boldsymbol{a}'\left(\Gamma^*_{t,T}-\Gamma_{t/T}\right)\boldsymbol{a}+\inf_t\inf_{\|\boldsymbol{a}\|=1}\boldsymbol{a}'\Gamma_{t/T}\boldsymbol{a}\geq 2\pi f_--\frac{d^{3/2}C}{T}\,.\end{aligned}$$

Then, for $T\geq T_0=Cd^{3/2}(\pi f_-)^{-1}$ we have $\|(\Gamma^*_{t,T})^{-1}\|\leq(\pi f_-)^{-1}$. Lemma 16 ensures that $\|\boldsymbol{\theta}_{t/T}\|\leq 2^d$ and the result follows with $C_1=Cd^{1/2}(\pi f_-)^{-1}(d2^d+1)$. □

Theorem 5.8.1. *Let $d\in\mathbb{N}^*,\beta\geq 2,R>0$, $f_-=0$ and $h:[0,1]\to\mathbb{R}$. Let $k\in\mathbb{N}$ and $\alpha\in(0,1]$ be uniquely defined such that $\beta=k+\alpha$ and consider $M\in 2\mathbb{N}^*$ with $M>d$. Suppose that Assumptions (**M**-3) and (**H**) hold. Then, for all $-d\leq j\leq d$ we have*

$$\mathbb{E}\left[\widehat{\gamma}_M\left(\frac{t}{T},j\right)\right]=\gamma\left(\frac{t}{T},j\right)+\sum_{\ell=1}^{k}c_{h,f,j,\ell}\left(\frac{M}{T}\right)^{\ell}+O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M}+\left(\frac{M}{T}\right)^{\beta}\right)\,,$$

where $c_{h,f,j,\ell}$ only depends on h, the spectral density f, j and ℓ. If $h(x)=h(1-x)$ for all $x\in[0,1]$, then $c_{h,f,j,1}=0$.

Our proof of Theorem 5.8.1 can be found in Section 5.8.1. It uses the following lemma.

Lemma 18. *Let $\beta > 0$ and $R > 0$. Consider $f : \mathbb{R} \to \mathbb{R}$, a function in $\Lambda_1(\beta, R)$ and $a \in \mathbb{R}$. Let $k \in \mathbb{N}$ and $\alpha \in (0, 1]$ be uniquely defined such that $\beta = k + \alpha$. The function f admits the representation*

$$f(x) = \sum_{\ell=0}^{k} \frac{f^{(\ell)}(a)}{\ell!} (x-a)^{\ell} + f_k(x) \ , \tag{5.8.4}$$

where $f_k(x) = O_{\beta,R}((x-a)^{\beta})$.

Proof. The expression (5.8.4) corresponds to the Taylor expansion of the function f. Without loss of generality, let $x > a$. We just need to proof that the remainder term is bounded by $(x-a)^{\beta}$ up to a constant. Using the definition of the norm $|\cdot|_{\beta}$ we have $f_k^{(k)}(x) \leq R(x-a)^{\alpha}$. The result follows by integrating k times the previous inequality. □

5

5.8.1 Proof of Theorem 5.8.1

Without loss of generality let $j \geq 0$. We start by expressing $\widehat{\gamma}_M$ in function of γ^*

$$\begin{aligned}
\mathbb{E}\left[\widehat{\gamma}_M\left(\frac{t}{T}, j\right)\right] &= \frac{1}{H_M} \sum_{\substack{t_1,t_2=1 \\ t_1-t_2=\ell}}^{M} h\left(\frac{t_1}{M}\right) h\left(\frac{t_2}{M}\right) \mathbb{E}\left[X_{t+t_1-M/2,T} X_{t+t_2-M/2,T}\right] \ , \\
&= \frac{1}{H_M} \sum_{s=j+1}^{M} h\left(\frac{s}{M}\right) h\left(\frac{s-j}{M}\right) \gamma^*\left(t+s-\frac{M}{2}, T, j\right) \ .
\end{aligned}$$

Since Inequality (5.2.3) guaranties that

$$\left|\gamma^*\left(t+s-\frac{M}{2}, T, j\right) - \gamma\left(\frac{t+s-M/2}{T}, j\right)\right| = O_R\left(\frac{1}{T}\right) \ , \tag{5.8.5}$$

we evaluate

$$\gamma_{M,j} = \frac{1}{H_M} \sum_{s=j+1}^{M} h\left(\frac{s}{M}\right) h\left(\frac{s-j}{M}\right) \gamma\left(\frac{t+s-M/2}{T}, T, j\right) \ ,$$

and then use the expression of $\gamma_{M,j}$ for computing $\mathbb{E}[\widehat{\gamma}_M(t/T, j)]$.
We apply Lemma 18 on the first component of f. The corresponding ℓ-th derivative is denoted by ∂_1^{ℓ}.

$$f\left(\frac{t-M/2+s}{T}, \lambda\right) = \sum_{\ell=0}^{k} \frac{\partial_1^{\ell} f(t/T, \lambda)}{\ell!} \left(\frac{-M/2+s}{T}\right)^{\ell} + f_k\left(\frac{t-M/2+s}{T}, \lambda\right) \ ,$$

with $f_k((t-M/2+s)/T,\lambda) = O_{\beta,R}((M/T)^\beta)$. Then

$$\gamma_{M,j} = \frac{1}{H_M}\int_{-\pi}^{\pi} \exp(ij\lambda) \sum_{s=j+1}^{M} h\left(\frac{s}{M}\right) h\left(\frac{s-j}{M}\right) f\left(\frac{t-M/2+s}{T},\lambda\right) \mathrm{d}\lambda =$$
$$\sum_{\ell=0}^{k}\int_{-\pi}^{\pi} \frac{\partial_1^\ell f(t/T,\lambda)}{\ell!} \exp(ij\lambda) \frac{1}{H_M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right) h\left(\frac{s-j}{M}\right)\left(\frac{-M/2+s}{T}\right)^\ell \mathrm{d}\lambda$$
$$+\int_{-\pi}^{\pi} \exp(ij\lambda) \frac{1}{H_M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right) h\left(\frac{s-j}{M}\right) f_k\left(\frac{t-M/2+s}{T},\lambda\right) \mathrm{d}\lambda\,. \quad (5.8.6)$$

Note that for all $\ell = 1,\dots,k$

$$\frac{1}{H_M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right) h\left(\frac{s-j}{M}\right)\left(\frac{-M/2+s}{T}\right)^\ell =$$
$$\left(\frac{M}{T}\right)^\ell \frac{M}{H_M}\frac{1}{M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right) h\left(\frac{s-j}{M}\right)\left(-\frac{1}{2}+\frac{s}{M}\right)^\ell\,. \quad (5.8.7)$$

Since h is piecewise C^1, maybe except for N values of s in $j+1,\dots,M$ we have

$$h\left(\frac{s-j}{M}\right) = h\left(\frac{s}{M}\right) + O_{\|h'\|_\infty}\left(\frac{d}{M}\right),$$

and we express the right-hand side of (5.8.7) as two right Riemann sums

$$\frac{1}{M}\sum_{s=j+1}^{M} h^2\left(\frac{s}{M}\right)\left(-\frac{1}{2}+\frac{s}{M}\right)^\ell = \int_0^1 h^2(u)\left(u-\frac{1}{2}\right)^\ell \mathrm{d}u + \frac{\|h\|_\infty\left(\|h'\|_\infty + \ell\,\|h\|_\infty\right)}{2^\ell M} o_{1,M,\ell}$$
$$+\frac{d\,\|h\|_\infty^2}{M} o_{2,M,\ell}\,,$$
$$\frac{1}{M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right)\left(-\frac{1}{2}+\frac{s}{M}\right)^\ell = \int_0^1 h(u)\left(u-\frac{1}{2}\right)^\ell \mathrm{d}u + \frac{\|h'\|_\infty + 2\ell\,\|h\|_\infty}{2^{\ell+1}M} o_{3,M,\ell} + \frac{d\,\|h\|_\infty}{M} o_{4,M,\ell}\,,$$

with $|o_{i,M,\ell}| \le 1$ for $i = 1,\dots,4$. Analogously

$$\frac{M}{H_M} = \left(\int_0^1 h^2(u)\,\mathrm{d}u\right)^{-1}\left(1 + \frac{\|h\|_\infty\,\|h'\|_\infty}{2^\ell M} o_{M,\ell}\right), \quad (5.8.8)$$

with $|o_{M,\ell}| \leq 1$. Hence

$$\frac{1}{H_M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right)h\left(\frac{s-j}{M}\right)\left(\frac{-M/2+s}{T}\right)^{\ell} = c_{h,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty}\left(\frac{1}{M}\right)\left(\frac{M}{T}\right)^{\ell} . \quad (5.8.9)$$

Observe that $c_{h,0} = 1$ and $c_{h,1} = 0$ if $h(x) = h(1-x)$ for all $x \in [0,1]$. Using (5.8.9) and the upper bound on f_k, we express the terms of the second and third lines of (5.8.6) as follows

$$\int_{-\pi}^{\pi} \frac{\partial_1^{\ell} f(t/T,\lambda)}{\ell!} \exp(ij\lambda)\frac{1}{H_M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right)h\left(\frac{s-j}{M}\right)\left(\frac{-M/2+s}{T}\right)^{\ell} \mathrm{d}\lambda = c_{h,f,j,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M}\right),$$

$$\int_{-\pi}^{\pi} \phi(\lambda)\frac{1}{H_M}\sum_{s=j+1}^{M} h\left(\frac{s}{M}\right)h\left(\frac{s-j}{M}\right) f_k\left(\frac{t-M/2+s}{T},\lambda\right)\mathrm{d}\lambda = O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\left(\frac{M}{T}\right)^{\beta}\right),$$

where, in particular $c_{h,f,j,0} = \gamma(t/T, j)$. This implies that

$$\gamma_{M,j} = \gamma\left(\frac{t}{T}, j\right) + \sum_{\ell=1}^{k} c_{h,f,j,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^{\beta}\right).$$

Note that the relation (5.8.5) together with (5.8.9) evaluated in $\ell = 0$ allow to conclude the proof.

5.9 Proof of bounds of the estimation risk

5.9.1 Proof of Theorem 5.5.1

We start by enunciating and proving the following.

Lemma 19. *Let d be a positive integer. Consider the $d \times d$ real non singular matrices Γ and $\widehat{\Gamma}$ and the vectors $\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}, \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ satisfying the relations*

$$\Gamma\boldsymbol{\theta} = \boldsymbol{\gamma}\,, \quad (5.9.1)$$

$$\widehat{\Gamma}\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\gamma}}\,. \quad (5.9.2)$$

Then, for any $k \in \mathbb{N}$ we have

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left(\Gamma^{-1} + \sum_{\ell=1}^{k}\left(\Gamma^{-1}\left(\Gamma - \widehat{\Gamma}\right)\right)^{\ell}\right)(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \sum_{\ell=1}^{k+1}\left(\Gamma^{-1}\left(\Gamma - \widehat{\Gamma}\right)\right)^{\ell}\boldsymbol{\theta} + \left(\Gamma^{-1}\left(\Gamma - \widehat{\Gamma}\right)\right)^{k+1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \quad (5.9.3)$$

Proof. From Equations (5.9.1) and (5.9.2) we get

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \;=\; \Gamma^{-1}\left[\left(\Gamma - \widehat{\Gamma}\right)\widehat{\boldsymbol{\theta}} + \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right] .$$

The result follows by applying recursion. □

Gathering together Assumption (**C**) and Theorem 5.8.1 yields

$$\widehat{\gamma}_M\left(\frac{t}{T}, j\right) = \gamma\left(\frac{t}{T}, j\right) + \sum_{\ell=1}^{k} c_{h,f,j,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^{\beta}\right) + u_M\left(\frac{t}{T}, j\right) , \quad (5.9.4)$$

where $u_M(t/T, j) = O_{L^\bullet,\|h\|_\infty,\|h'\|_\infty,R,j}(M^{-1/2})$ and $c_{h,f,j,1} = 0$ if $h(x) = h(1-x)$ for all $x \in [0,1]$.

For the sake of simplicity, we drop t, T in the notation and set $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{t/T}$, $\widehat{\boldsymbol{\gamma}}_M = \widehat{\boldsymbol{\gamma}}_{t,T,M}$, $\Gamma \equiv \Gamma_{t/T}$ and $\widehat{\Gamma}_M \equiv \widehat{\Gamma}_{t,T,M}$. Using the expression (5.9.4), we choose $j = 0, \ldots, d$ and obtain

$$\Gamma - \widehat{\Gamma}_M = \sum_{\ell=1}^{k} C_{h,f,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^{\beta}\right) + U_M , \quad (5.9.5)$$

$$\widehat{\boldsymbol{\gamma}}_M - \boldsymbol{\gamma} = \sum_{\ell=1}^{k} \boldsymbol{c}_{h,f,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^{\beta}\right) + \boldsymbol{u}_M , \quad (5.9.6)$$

where the matrices $C_{h,f,\ell} \in \mathbb{R}^{d\times d}$ and the vectors $\boldsymbol{c}_{h,f,\ell} \in \mathbb{R}^d$ only depend on h, f and ℓ. Furthermore $U_M = O_{L^\bullet,d,\|h\|_\infty,\|h'\|_\infty,\beta,R}(M^{-1/2})$ and $\boldsymbol{u}_M = O_{L^\bullet,d,\|h\|_\infty,\|h'\|_\infty,\beta,R}(M^{-1/2})$. Again $C_{h,f,1} = 0$ and $\boldsymbol{c}_{h,f,1} = 0$ if $h(x) = h(1-x)$ for all $x \in [0,1]$.

Note that the product of $q > 0$ expressions for the form

$$\sum_{\ell=1}^{k} C_{h,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^{\beta}\right) + V_M ,$$

with $V_M = O_{L^\bullet,d,\|h\|_\infty,\|h'\|_\infty,\beta,R}(M^{-1/2})$, has the form

$$\sum_{\ell=q}^{k} D_{h,\ell}\left(\frac{M}{T}\right)^{\ell} + O_{d,\|h\|_\infty,\|h'\|_\infty,\beta,R}\left(\frac{1}{M} + \left(\frac{M}{T}\right)^{\beta}\right) + W_M ,$$

with $W_M = O_{L^\bullet,d,\|h\|_\infty,\|h'\|_\infty,\beta,R}(M^{-1/2})$ and $D_{h,\ell} = 0$ for $\ell \in [q, 2q)$ if all the $C_{h,1}$ of the factors vanish. This remark, together with (5.9.3) of Lemma 19, (5.9.5), (5.9.6), and the bounds provided by Lemma 15, Lemma 16 and Lemma 22 imply what is claimed in (5.5.2).

5.9.2 Proof of Theorem 5.5.2

For each $\ell = 0, \ldots, k$ plug $M/2^\ell$ instead of M into Equation (5.5.2), multiply the resulting expression by α_ℓ and sum. Matrix A (definition below Equation (5.5.3)) is a non singular Vandermonde matrix and $\boldsymbol{\alpha}$ is well defined. If $h(x) = h(1-x)$ for $x \in [0,1]$ we can remove the second row of matrix A because the first order term of (5.5.2) is zero.

5.10 Useful results on time varying autoregressive processes

Let δ be a positive real number. Consider the set

$$s_{(p)}(\delta) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^p : \boldsymbol{\theta}(z) = 1 - \sum_{k=1}^{p} \theta_k z^k \neq 0, \text{ for } |z| < \delta^{-1} \right\} . \tag{5.10.1}$$

As an immediate consequence of Hurwitz's theorem (see (Conway, 1973, Theorem 2.5) or (Gamelin, 2001, Section 3, Chapter VIII)) we obtain the following lemma.

Lemma 20. *For any $\delta > 0$ the set $s_{(p)}(\delta)$ defined by Equation (5.10.1) is closed and $\min_{\boldsymbol{\theta} \in s_{(p)}(\delta)} \|\boldsymbol{\theta}\|_\infty > 0$ only depends on p and δ.*

Since $\boldsymbol{\theta} \in s_p(\delta)$ and $\sigma \in \Lambda_1(\beta, R)$, thank to Lemma 20 we have that for any $\lambda \in \mathbb{R}$ the spectral density $f(\cdot, \lambda)$ belongs to a $\Lambda_1(\beta, R')$ with R' depending only on R, δ and continuously on β. A direct consequence of Lemma 20 is given below.

Lemma 21. *Let $\delta \in (0,1), \beta > 0, R > 0$ and $\rho \in [0,1]$. Suppose that Assumptions (M-4) and (I) hold. There exist two constants f_-, f_+, depending only on p, δ, ρ and σ_+ such that $0 < f_- \leq f(u, \lambda) \leq f_+$ for all $u, \lambda \in \mathbb{R}$.*

5.11 Useful results on weakly stationary processes

In the context of real weakly stationary processes (see Brockwell and Davis (2002) and Shumway and Stoffer (2011)), the autocovariance matrix of $(X_t)_{t \in \mathbb{Z}}$, that we call Γ_d, is Toeplitz and symmetric. Observe that

$$\Gamma_d = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(d-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(d-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(d-1) & \gamma(d-2) & \gamma(d-3) & \dots & \gamma(0) \end{bmatrix} . \tag{5.11.1}$$

Proposition 3. *A complex-valued function defined on $\mathbb{Z}$ is the autocovariance function of a weakly stationary process $(X_t)_{t \in \mathbb{Z}}$ taking values in $\mathbb{C}$ if and only if the following two properties hold.*

(i) Hermitian symmetry: for all $s \in \mathbb{Z}$,

$$\gamma(-s) = \overline{\gamma(s)} .$$

(ii) Nonnegativity: for all $d \in \mathbb{N}^*$ and $a_1, \dots, a_d \in \mathbb{C}$,

$$\sum_{i=1}^{d} \sum_{j=1}^{d} \overline{a_i} \gamma\left(t_i - t_j\right) a_j \geq 0 .$$

A crucial concept in the study of weakly stationary processes is the spectral measure, defined from the autocovariance function γ (see (Brockwell and Davis, 2002, Theorem 4.3.1)). We denote by $\mathcal{B}([-\pi,\pi))$ the Borel σ-algebra associated with $[-\pi,\pi)$.

Theorem 5.11.1 (Herglotz)**.** *A sequence γ is nonnegative definite and hermitian in the sense of Proposition 3 if and only if there exists a finite nonnegative measure ν on $([-\pi,\pi),\mathcal{B}([-\pi,\pi)))$ such that, for all $s\in\mathbb{Z}$:*

$$\gamma(s)=\int_{-\pi}^{\pi}\exp(ik\lambda)\,\nu(\mathrm{d}\lambda)\ . \tag{5.11.2}$$

Furthermore, the measure ν is unique.

The next result links the spectral density function (when it exists) and the spectrum of the covariance matrix Γ_d.

Lemma 22. *Assume that the autocovariance function γ has a spectral density function $f\in[f_-,f_+]$ with $f_-\leq f_+$. For any $d\in\mathbb{N}^*$, the spectrum of the covariance matrix (Equation (5.11.1)) is contained in $[2\pi f_-,2\pi f_+]$.*

Proof. Consider $\boldsymbol{a}=[a_1\ \ldots\ a_d]'\in\mathbb{R}^d$. If we express γ using the representation (5.11.2) we obtain

$$\boldsymbol{a}'\Gamma_d\boldsymbol{a}=\int_{-\pi}^{\pi}\left|\sum_{j=1}^{d}a_j\exp(ij\lambda)\right|^2 f(\lambda)\,\mathrm{d}\lambda\ .$$

Therefore $2\pi f_+\sum_{j=1}^d a_j^2\geq\boldsymbol{a}'\Gamma_d\boldsymbol{a}\geq 2\pi f_-\sum_{j=1}^d a_j^2$. Choosing $\boldsymbol{a}$ as any eigenvector of Γ_d the result follows. □

The lemma below is similar in flavor to the statistical result of (Whittle, 1963, Section 3). It is also a classical property of orthogonal polynomials (see (Grenander and Szegő, 1984, Section 2.4)). We provide an elementary proof.

Lemma 23. *Let γ be a real autocovariance function (in the sense of Proposition 3) such that for any $d\in\mathbb{N}^*$, the covariance matrix Γ_d defined by Equation (5.11.1) is positive-definite. Denote the vector $\boldsymbol{\gamma}_d=[\gamma(1)\ \ldots\ \gamma(d)]'$ and let $\widehat{\boldsymbol{\theta}}=[\widehat{\theta}_1\ \ldots\ \widehat{\theta}_d]'=\Gamma_d^{-1}\boldsymbol{\gamma}_d$. Then, all the roots of the polynomial $\widehat{\boldsymbol{\theta}}(z)=1-\sum_{j=1}^d\widehat{\theta}_j z^j$ are in the set $\{z\in\mathbb{C}:|z|>1\}$.*

Proof. For $j=1,\ldots,d$, let $\boldsymbol{e}_j=[0\ \ldots\ 1\ \ldots\ 0]'$ be the $\mathbb{R}^d$- vector having a 1 in the j-th position and zero everywhere else. Consider also the matrix

$$A=\begin{bmatrix}\widehat{\theta}_1 & \widehat{\theta}_2 & \ldots & \ldots & \widehat{\theta}_d\\ 1 & 0 & \ldots & \ldots & 0\\ 0 & 1 & 0 & \ddots & 0\\ \vdots & 0 & \ddots & \ddots & \vdots\\ 0 & \ldots & 0 & 1 & 0\end{bmatrix}=\begin{bmatrix}\widehat{\boldsymbol{\theta}}'\\ \boldsymbol{e}_1'\\ \vdots\\ \boldsymbol{e}_{d-1}'\end{bmatrix}\ .$$

Since the roots of $\widehat{\boldsymbol{\theta}}(z)$ are the inverses of the eigenvalues of A, we need to proof that these eigenvalues are inside the open unit disk. Observe that

$$\Gamma_d - A\Gamma_d A' = \Gamma_d - \begin{bmatrix} \widehat{\boldsymbol{\theta}}'\Gamma_d\widehat{\boldsymbol{\theta}} & \widehat{\boldsymbol{\theta}}'\Gamma_d\boldsymbol{e}_1 & \dots & \dots & \widehat{\boldsymbol{\theta}}'\Gamma_d\boldsymbol{e}_{d-1} \\ \boldsymbol{e}_1'\Gamma_d\widehat{\boldsymbol{\theta}} & \boldsymbol{e}_1'\Gamma_d\boldsymbol{e}_1 & \dots & \dots & \boldsymbol{e}_1'\Gamma_d\boldsymbol{e}_{d-1} \\ \boldsymbol{e}_2'\Gamma_d\widehat{\boldsymbol{\theta}} & \boldsymbol{e}_2'\Gamma_d\boldsymbol{e}_1 & \dots & \dots & \boldsymbol{e}_2'\Gamma_d\boldsymbol{e}_{d-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{e}_{d-1}'\Gamma_d\widehat{\boldsymbol{\theta}} & \boldsymbol{e}_{d-1}'\Gamma_d\boldsymbol{e}_1 & \dots & \dots & \boldsymbol{e}_{d-1}'\Gamma_d\boldsymbol{e}_{d-1} \end{bmatrix}.$$

Because Γ_d is a Toeplitz matrix, its (i, j)-th entries, and those of $A\Gamma_d A'$ are equal for $i, j \geq 2$. The definition of $\widehat{\boldsymbol{\theta}}$ implies also the equality of the (i, j)-th entries of both matrices when $i = 1, j \geq 2$ and $i \geq 2, j = 1$. Since $\widehat{\boldsymbol{\theta}}$ is the solution of $\Gamma_d\widehat{\boldsymbol{\theta}} = \boldsymbol{\gamma}_d$, we have that $\widehat{\theta}_j = -\Gamma_{d+1,d,j}/\det(\Gamma_d)$ where $\Gamma_{d,i,j}$ is the cofactor of the (i, j)-th entry of Γ_d. Finally, in the position $(1, 1)$ we have $\gamma(0) - \widehat{\boldsymbol{\theta}}'\boldsymbol{\gamma}_d = \sum_{j=0}^{d}\gamma(j)\Gamma_{d+1,d,j}/\det(\Gamma_d) = \det(\Gamma_{d+1})/\det(\Gamma_d) > 0$. Consider now λ, an eigenvalue of A and the corresponding eigenvector $\boldsymbol{v} \neq 0$. We verify that $\boldsymbol{v} = [\lambda^{d-1} \ \dots \ \lambda \ 1]'v_d$. From the previous analysis we get $\bar{\boldsymbol{v}}'(\Gamma_d - A\Gamma_d A')\boldsymbol{v} = |\lambda|^{2d-2}(1 - |\lambda|^2)\det(\Gamma_{d+1})/\det(\Gamma_d)|v_d|^2 \geq 0$ and the inequality is strict if $\lambda \neq 0$. As claimed $|\lambda| < 1$. □

ACKNOWLEDGEMENTS

This work has been partially supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île de France (RDM-IdF) for the period 2012 - 2015 and by the Labex LMH (ANR-11-IDEX-003-02).

Bibliography

Bibliography

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, 21:243–247. 9, 10, 41, 42

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest. 10, 42

Akaike, H. (1978). Time series analysis and control through parametric models. In Findley, D., editor, *Applied Time Series Analysis*. Academic Press, New York. 10, 42

Alquier, P. and Li, X. (2012). Prediction of quantiles by statistical learning and application to gdp forecasting. In Ganascia, J.-G., Lenca, P., and Petit, J.-M., editors, *Discovery Science*, volume 7569 of *Lecture Notes in Computer Science*, pages 22–36. Springer Berlin Heidelberg. 63, 66

Alquier, P. and Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913. 4, 24, 36, 55, 63, 64, 66, 76, 77, 78, 84

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588. 18, 49

Anava, O., Hazan, E., Mannor, S., and Shamir, O. (2013). Online learning for time series prediction. *J. Mach. Learn. Res.*, 30:172–184. 84, 133

Andrieu, C. and Doucet, A. (1999). An improved method for uniform simulation of stable minimum phase real arma (p,q) processes. *Signal Processing Letters, IEEE*, 6(6):142–144. 75

Arkoun, O. (2011). Sequential adaptive estimators in nonparametric autoregressive models. *Sequential Anal.*, 30(2):229–247. 100

Atchadé, Y. F. (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.*, 8(2):235–254. 76

Audibert, J.-Y. (2004). *PAC-Bayesian Statistical Learning Theory*. PhD thesis, Université Pierre et Marie Curie-Paris VI. 22, 53, 63

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646. 23, 27, 54, 58, 84, 95

Audibert, J.-Y. and Catoni, O. (2010). Robust linear regression through pac-bayesian truncation. *arXiv preprint arXiv:1010.0072*. 8, 40

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794. 8, 40

Auer, P., Cesa-Bianchi, N., and Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms. *J. Comput. System Sci.*, 64(1):48–75. Special issue on COLT 2000 (Palo Alto, CA). 20, 51

Baranger, J. and Brezinski, C. (1991). *Analyse numérique*, volume 38. Hermann, Paris. Collection Méthodes. 135

Barron, A. R. (1987). Are bayes rules consistent in information? In Cover, T. M. and Gopinath, B., editors, *Open Problems in Communication and Computation*, pages 85–91. Springer New York. 18, 49, 133

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054. 17, 48

Baxendale, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.*, 15(1B):700–738. 18, 49

Beadle, E. R. and Djurić, P. M. (1999). Uniform random parameter generation of stable minimum-phase real arma (p,q) processes. *Signal Processing Letters, IEEE*, 4(9):259–261. 75

Berk, K. N. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.*, 2:489–502. Collection of articles dedicated to Jerzy Neyman on his 80th birthday. 9, 41

Bhansali, R. J. (1978). Linear prediction by autoregressive model fitting in the time domain. *Ann. Statist.*, 6(1):224–231. 9, 10, 41

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237. 16, 47

Birgé, L. and Massart, P. (2000). An adaptive compression algorithm in Besov spaces. *Constr. Approx.*, 16(1):1–36. 17, 48

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142. 12, 44

Bottou, L. (2012). Large-scale machine learning with stochastic gradient descent. In *Statistical learning and data science*, Comput. Sci. Data Anal. Ser., pages 17–25. CRC Press, Boca Raton, FL. 12, 44

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140. 18, 49

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer-Verlag, New York, second edition. With 1 CD-ROM (Windows). 4, 10, 36, 42, 129, 146, 147

Brockwell, P. J. and Davis, R. A. (2006). *Time series: theory and methods*. Springer Series in Statistics. Springer, New York. Reprint of the second (1991) edition. 65, 85, 86, 141

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat. 17, 48

Catoni, O. (1997). A mixture approach to universal model selection. Technical report, École Normale Supérieure. 18, 49, 84, 133

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. 18, 20, 22, 26, 49, 51, 53, 57, 63, 64, 76, 88, 93, 100, 104, 133

Čencov, N. N. (1962). A bound for an unknown distribution density in terms of the observations. *Dokl. Akad. Nauk SSSR*, 147:45–48. 16, 47

Cesa-Bianchi, N. (1999). Analysis of two gradient-based algorithms for on-line regression. *J. Comput. System Sci.*, 59(3):392–411. 13, 44

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press, Cambridge. 20, 51, 63, 84, 88, 89, 95, 133

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. (2005). Improved second-order bounds for prediction with expert advice. In *Learning theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 217–232. Springer, Berlin. 94

Conway, J. B. (1973). *Functions of one complex variable*. Springer-Verlag, New York-Heidelberg. Graduate Texts in Mathematics, 11. 146

Coulon-Prieur, C. and Doukhan, P. (2000). A triangular central limit theorem under a new weak dependence condition. *Statist. Probab. Lett.*, 47(1):61–68. 66

Cover, T. M. (1991). Universal portfolios. *Math. Finance*, 1(1):1–29. 71

Dahlhaus, R. (1996a). Maximum likelihood estimation and model selection for locally stationary processes. *J. Nonparametr. Statist.*, 6(2-3):171–191. 130

Dahlhaus, R. (1996b). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.*, 62(1):139–168. 5, 6, 36, 37, 38, 86, 96, 98, 128, 129, 130, 137

Dahlhaus, R. (2009). Local inference for locally stationary time series based on the empirical spectral measure. *J. Econometrics*, 151(2):101–112. 5, 37, 84, 86, 87, 113

Dahlhaus, R. (2012). Locally stationary processes. In Rao, T., Rao, S., and Rao, C., editors, *Time Series Analysis: Methods and Applications*, volume 30 of *Handbook of Statistics*, pages 351–413. North Holland. 128

Dahlhaus, R. and Giraitis, L. (1998). On the optimal segment length for parameter estimates for locally stationary time series. *J. Time Ser. Anal.*, 19(6):629–655. 6, 30, 38, 60, 113, 128, 131, 134, 135, 139, 140

Dahlhaus, R. and Polonik, W. (2006). Nonparametric quasi-maximum likelihood estimation for Gaussian locally stationary processes. *Ann. Statist.*, 34(6):2790–2824. 87

Dahlhaus, R. and Polonik, W. (2009). Empirical spectral processes for locally stationary time series. *Bernoulli*, 15(1):1–39. 87, 98

Dahlhaus, R. and Subba Rao, S. (2006). Statistical inference for time-varying ARCH processes. *Ann. Statist.*, 34(3):1075–1114. 6, 38

Dalalyan, A. S. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61. 22, 53, 63, 84

Dalalyan, A. S. and Tsybakov, A. B. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443. 17, 48, 76

Dedecker, J., Doukhan, P., Lang, G., León R., J. R., Louhichi, S., and Prieur, C. (2007). *Weak dependence: with examples and applications*, volume 190 of *Lecture Notes in Statistics*. Springer, New York. 2, 3, 34, 35, 64

Dedecker, J. and Prieur, C. (2005). New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, 132(2):203–236. 64

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921. 15, 17, 46, 48

Doukhan, P. (2003). Models, inequalities, and limit theorems for stationary sequences. In *Theory and applications of long-range dependence*, pages 43–100. Birkhäuser Boston, Boston, MA. 3, 35

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.*, 84(2):313–342. 2, 3, 33, 35

Doukhan, P. and Wintenberger, O. (2008). Weakly dependent chains with infinite memory. *Stochastic Process. Appl.*, 118(11):1997–2013. 90

Duflo, M. (1997). *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the 1990 French original by Stephen S. Wilson and revised by the author. 127

Efroimovich, S. Y. and Pinsker, M. (1984). A self-educating nonparametric filtration algorithm. *Automation and Remote Control*, 45:58–65. 17, 48

Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.*, 43:170–180. 16, 47

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, 38(2):1034–1070. 18, 49

Foster, D. P. (1991). Prediction in the worst case. *Ann. Statist.*, 19(2):1084–1090. 20, 51

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.*, 121(2):256–285. 18, 49

Gamelin, T. W. (2001). *Complex analysis*. Undergraduate Texts in Mathematics. Springer-Verlag, New York. 146

Gerchinovitz, S. (2011). *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques*. PhD thesis, Université Paris Sud-Paris XI. 84, 88

Gerchinovitz, S. (2013). Sparsity regret bounds for individual sequences in online linear regression. *J. Mach. Learn. Res.*, 14:729–769. 20, 51

Geyer, C. J. (1992). *Practical Markov chain Monte Carlo*. JSTOR. 18, 49

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263. 125

Gill, R. D. and Levit, B. Y. (1995). Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79. 16, 47

Giraud, C. (2015). *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. 18, 49

Granger, C. W. J. (1964). *Spectral analysis of economic time series*. In association with M. Hatanaka. Princeton Studies in Mathematical Economics, No. I. Princeton University Press, Princeton, N.J. 5, 36

Grenander, U. and Szegő, G. (1984). *Toeplitz forms and their applications*. Chelsea Publishing Co., New York, second edition. 147

Grenier, Y. (1983). Time-dependent ARMA modeling of nonstationary signals. *IEEE Transactions on ASSP*, 31(4):899–911. 86

Haussler, D., Kivinen, J., and Warmuth, M. K. (1998). Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inform. Theory*, 44(5):1906–1925. 18, 49, 95

Hsu, D., Kakade, S. M., and Zhang, T. (2011). An analysis of random design linear regression. In *Proc. COLT*. Citeseer. 8, 40

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307. 10, 42

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probab. Surv.*, 1:299–320. 18, 49

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.*, 16(4):312–334. 18, 49

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712. 18, 23, 49, 54, 94, 95, 133

Kalai, A. and Vempala, S. (2002). Efficient algorithms for universal portfolios. *J. Mach. Learn. Res.*, 3(Spec. Issue Comput. Learn. Theory):423–440. 71

Künsch, H. R. (1995). A note on causal solutions for locally stationary ar-processes. Preprint ETH Zürich. 65, 86, 98

Łatuszyński, K., Miasojedow, B., and Niemiro, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, 19(5A):2033–2066. 64

Łatuszyński, K. and Niemiro, W. (2011). Rigorous confidence bounds for MCMC under a geometric drift condition. *J. Complexity*, 27(1):23–38. 17, 18, 48, 49, 64, 70, 71, 82

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470. 100

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659. 17, 48

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410. 18, 49, 63, 84, 133

Lewis, R. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting. *J. Multivariate Anal.*, 16(3):393–411. 9, 41

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Inform. and Comput.*, 108(2):212–261. 18, 49, 133

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580. 138

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. 4, 36, 120

McAllester, D. A. (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, pages 164–170 (electronic). ACM, New York. 22, 53

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121. 70

Meyn, S. and Tweedie, R. L. (2009). *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition. With a prologue by Peter W. Glynn. 17, 18, 48, 49

Moulines, E., Priouret, P., and Roueff, F. (2005). On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654. 6, 13, 31, 38, 44, 61, 73, 84, 97, 100, 110, 112, 113, 114, 115, 116, 128, 132, 133, 135, 137

Nemirovskiĭ, A. S. (1990). Necessary conditions for efficient estimation of functionals of a nonparametric signal observed in white noise. *Teor. Veroyatnost. i Primenen.*, 35(1):83–91. 16, 47

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes.(With discussion). *J. Roy. Statist. Soc. Ser. B*, 27:204–237. 5, 36, 128

Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575. 84

Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908. 4, 36, 64, 77

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71. 18, 49, 70

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A.*, 42:43–47. 2, 34

Rosenblatt, M. (1980). Linear processes and bispectra. *J. Appl. Probab.*, 17(1):265–270. 3, 35

Sancetta, A. (2010). Recursive forecast combination for dependent heterogeneous data. *Econometric Theory*, 26(2):598–631. 94

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464. 10, 42

Shumway, R. H. and Stoffer, D. S. (2011). *Time series analysis and its applications*. Springer Texts in Statistics. Springer, New York, third edition. With R examples. 146

Stoltz, G. (2011). Contributions to the sequential prediction of arbitrary sequences: applications to the theory of repeated games and empirical studies of the performance of the aggregation of experts. Habilitation à diriger des recherches, Université Paris Sud-Paris XI. 20, 26, 51, 57, 84, 88, 103

Subba Rao, S. (2006). On some nonstationary, nonlinear random processes and their stationary approximations. *Adv. in Appl. Probab.*, 38(4):1155–1172. 6, 38

Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society, Series B*, 42(3):245–292. 7, 39, 87

Tsybakov, A. B. (2003). Optimal rates of aggregation. In Schölkopf, B. and Warmuth, M. K., editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 303–313. Springer Berlin Heidelberg. 23, 27, 54, 58, 94, 95

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. 15, 16, 46, 47, 99, 108

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142. 22, 52

Volkonskiĭ, V. A. and Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theor. Probability Appl.*, 4:178–197. 2, 34

Vovk, V. (1998). A game of prediction with expert advice. *J. Comput. System Sci.*, 56(2):153–173. Eighth Annual Workshop on Computational Learning Theory (COLT) (Santa Cruz, CA, 1995). 95

Vovk, V. (2006). On-line regression competitive with reproducing kernel Hilbert spaces (extended abstract). In *Theory and applications of models of computation*, volume 3959 of *Lecture Notes in Comput. Sci.*, pages 452–463. Springer, Berlin. 20, 51

Vovk, V. G. (1990). Aggregating strategies. In *Proc. Third Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA. Morgan Kaufmann. 18, 49, 84, 133

Wang, Z., Paterlini, S., Gao, F., and Yang, Y. (2014). Adaptive minimax regression estimation over sparse ℓ_q-hulls. *Journal of Machine Learning Research*, 15:1675–1711. 94

Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50:129–134. 147

Yang, Y. (2000a). Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161. 17, 18, 48, 49, 84, 94, 100, 133

Yang, Y. (2000b). Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87. 100

Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20(1):176–222. 18, 22, 23, 27, 49, 53, 54, 58, 84, 93, 94, 95, 124, 133

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599. 16, 47

Aggregation of time series predictors, optimality in a locally stationary context

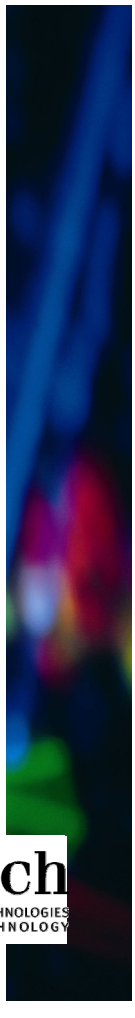
Andrés Sanchez Pérez

RESUME : Cette thèse regroupe nos résultats sur la prédiction de séries temporelles dépendantes. Le document comporte trois chapitres principaux où nous abordons des problèmes différents. Le premier concerne l'agrégation de prédicteurs de décalages de Bernoulli causales, en adoptant une approche Bayésienne. Le deuxième traite de l'agrégation de prédicteurs de ce que nous définissions comme processus sous-linéaires. Une attention particulaire est portée aux processus autorégressifs localement stationnaires variables dans le temps, nous examinons un schéma de prédiction adaptative pour eux. Dans le dernier chapitre nous étudions le modèle de régression linéaire pour une classe générale de processus localement stationnaires.

MOTS-CLEFS : séries temporelles non stationnaires, Causal Bernoulli Shifts, processus autorégressifs variables dans le temps, agrégation à poids exponentiels, apprentissage en ligne, prédiction adaptative.

ABSTRACT : This thesis regroups our results on dependent time series prediction. The work is divided into three main chapters where we tackle different problems. The first one is the aggregation of predictors of Causal Bernoulli Shifts using a Bayesian approach. The second one is the aggregation of predictors of what we define as sub-linear processes. Locally stationary time varying autoregressive processes receive a particular attention ; we investigate an adaptive prediction scheme for them. In the last main chapter we study the linear regression problem for a general class of locally stationary processes.

KEY-WORDS : non-stationary time series, Causal Bernoulli Shift, time varying autoregressive processes, exponential weighted aggregation, online learning, adaptive prediction.



TELECOM ParisTech

EDiTE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY